Structure and Function Studies of Clostridium difficile Surface-Associated Proteins

Bradshaw, Will

Award date:
2017

Awarding institution:
University of Bath

Structure and Function Studies of *Clostridium difficile* Surface-Associated Proteins

William Bradshaw

A thesis submitted for the degree of Doctor of Philosophy

University of Bath
Department of Biology and Biochemistry

September 2017

Acknowledgements

Firstly, I would like to thank my supervisors Ravi Acharya and Cliff Shone for giving me the opportunity to study for a PhD and for their invaluable advice, support and guidance. Clearly, without them, this work would not have been possible. This thanks should also be extended to the University of Bath and Public Heath England for funding this work.

I would also like to thank everyone in Lab 0.34 at the University of Bath (25 of you in my years here) and other staff and PhD students at Bath for the many constructive and useful discussions we've had and the rabbit holes that they led to (and maybe the not so constructive or useful ones too!) and for everything that I've learnt from so many of you. Thanks too, go to everyone in the toxins group at PHE for their help and making me feel welcome during the time that I spent there. While I'd like to thank many from Bath and PHE individually, each time I start, I'm unable to whittle down the list of important contributions. Special thanks, though, should be given to Gyles Cozier for proof reading this thesis and to Jon Kirby, Abi Davies, and Chris Chambers for their groundwork, advice and the donation of plasmids.

Thanks should also be given to Diamond Light Source and the staff on MX beamlines and B21 for assistance with data collection and for their rapid response whenever equipment went wrong at times when any sensible person would be asleep.

Thanks go to Albert Bolhuis, Tony James, Steve Flower, Andrew Watt and Matthew Lloyd at the University of Bath for the donation of carbohydrates used in Cwp19 assays and to Matthew Lloyd for advice relating to these assays.

This research made use of the Balena High Performance Computing (HPC) Service at the University of Bath, without which, much of the data processing would have taken exponentially longer.

Ultimately, this work would not have been possible without the years of education from countless teachers and lecturers. Many of their lessons may seem like a lifetime ago but they all made contributions towards my development as a scientist.

Lastly, I wish to thank my parents, family and friends for their unwavering support, care and encouragement throughout my PhD and for many years before it, despite the inevitable comments of “I understand some of the words in the title”.

Thank you.

Abstract

The Gram-positive bacterium *Clostridium difficile* is the primary causative agent of antibiotic-associated diarrhoea and kills tens of thousands of people around the world every year. While a significant proportion of research into *C. difficile* has focused on the toxins produced by the bacterium, which are responsible for disease aetiology, effective methods of prevention and treatment are likely to result from research into a range of aspects of the bacterium.

One such aspect is the paracrystalline layer of protein found on the cell surface of a wide range of bacterial species and virtually all archaea known as an S-layer. The S-layer of *C. difficile* is mostly formed by the low- and high-molecular weight S-layer proteins LMW SLP and HMW SLP. HMW SLP possesses three cell wall binding domains. The *C. difficile* genome contains 28 other genes that also code for proteins with three cell wall binding domains, many of these proteins also possess other domains that confer a specific function on the protein.

In this thesis work resulting in the determination of high-resolution structures of the "functional" regions of three proteins coded for by these genes is described: Cwp84, Cwp19 and Cwp2. This work required use of experimental phasing methods and advanced molecular replacement techniques. The structures reveal information on the role of these proteins in the formation of the S-layer, carbohydrate metabolism and host cell adhesion, respectively. This adds to the growing body of knowledge on the S-layer of *C. difficile* and may, in the future, contribute to the development of novel therapeutics against the bacterium.

Contents

List of figures and tables

List of abbreviations

AFM	Atomic force microscopy
AP	Anionic polymer
APD	Autoprotease domain
ARI	Art Robbins Instruments
CAZy	Carbohydrate Active enZyme database
CCD	Charge coupled device
CDI	*C. difficile* infection
CROPS	Combined repetitive oligopeptides
Cryo-EM	Cryo electron microscopy
CWB2/CWBD	Cell wall binding 2/Cell wall binding domain
Cwp	Cell/Clostridial wall protein
DSSP	Dictionary of protein secondary structure
ELISA	Enzyme linked immunosorbent assay
GH	Glycoside hydrolase
GHL10	Glycoside hydrolase-like 10
GSH	Glutathione
GTD	Glucosyltransferase domain
H&L	Heavy and light (molecular dimensions screen)
H/L complex	Complex formed by HMW SLP and LMW SLP
HMW SLP	High-molecular weight S-layer protein
IP	Image plate
IPTG	Isopropyl-β-D-1-thiogalactopyranoside
LB	Lysogeny broth
LCT	Large clostridial toxin

LLG	Log likelihood gain
LMW SLP	Low-molecular weight S-layer protein
M1	Morpheus (Molecular Dimensions Screen)
M2	Morpheus II (Molecular Dimensions Screen)
MAD	Multi-wavelength anomalous diffraction
MIC	Minimum inhibitory concentration
MIR	Multiple isomorphous replacement
NAG	N-acetylglucosamine
NAM	N-acetylmuramic acid
NMR	Nuclear magnetic resonance
PAD	Pixel-array detector
PBL	Prosegment binding loop
PCT	Pre-Crystallisation Test
PDB	Protein data bank
PPI	Proton pump inhibitors
RF	Rotation function
RFZ	Rotation function Z-score
SAD	Single-wavelength anomalous diffraction
SAXS	Small angle X-ray scattering
SANS	Small angle neutron scattering
SEC	Size exclusion chromatography
SD	Standard deviation
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
SIR	Single isomorphous replacement
SH3	Src homology 3
S-layer	Surface layer

SLH	S-layer homology
SLP	S-layer protein
SRP	Signal recognition particle
TB	Terrific broth
TF	Translation function
TFZ	Translation function Z-score

Declaration of work by other parties

Figure 1.1A and 1.1B © Public Health England, donated by Cliff Shone

Figure 1.1C © American College of Gastroenterology

Figure 1.1D reproduced from Carrion *et al.* (2010)

Figure 1.1E reproduced from Chumbler *et al.* (2016)

Figure 1.5A and 1.5B reproduced from Cerquetti *et al.* (2000)

Figure 1.5C and 1.5D reproduced from Fagan *et al.* (2009)

Figure 1.5E and 1.5F reproduced from Kirby (2011)

Figure 2.1 produced by Wikimedia Commons users DrBob and Stannered, used under the Creative Commons Attribution-Share Alike 3.0 Unported licence.

Figure 7.2 reproduced from Usenik *et al.* (2017)

All DNA constructs were produced and donated by Jon Kirby or Chris Chambers at Public Health England.

Chapter 1

The S-layer of *Clostridium difficile*

1.1 Introduction

The bacterium most commonly known as *Clostridium difficile*, which has recently been proposed for reclassification as *Peptoclostridium difficile* (Yutin & Galperin, 2013), and as *Clostridioides difficile* (Lawson *et al.*, 2016) is the primary causative agent of antibiotic-associated diarrhoea (McFarland *et al.*, 2016). *C. difficile* is a rod-shaped, obligate-anaerobic, gram-positive, bacterium (figure 1.1A and B) that was first identified in 1935 (Hall, 1935) and is present in the healthy gut flora of around 5% of adults and 50% of infants (Guarner & Malagelada, 2003; Kachrimanidou & Malisiovas, 2011). As with all *Clostridia*, *C. difficile* is able to form spores that are particularly persistent, showing resistance to immune systems, antibiotics, and disinfectants, (Freeman & Wilcox, 2003; Gessler & Bohnel, 2006; Barra-Carrasco & Paredes-Sabja, 2014). The bacterium is usually nosocomially acquired and does not cause health issues due to competition from other micro-organisms, a phenomenon known as the "barrier effect" (Guarner & Malagelada, 2003).

C. difficile becomes pathogenic after disruption of the gut flora, primarily through the use of antibiotics - clindamycin, broad-spectrum penicillins, cephalosporins and fluoroquinolones are particularly prolific causative agents, although many antibiotics have been implicated (Viswanathan *et al.*, 2010). Susceptibility to *C. difficile* infection has also been shown to be increased in the elderly, patients being treated with immunosuppressants or proton-pump inhibitors (PPIs), or patients suffering from gastrointestinal diseases (Kachrimanidou & Malisiovas, 2011).

C. difficile infection (CDI) can result in mild to severe diarrhoea, colitis, pseudomembranous colitis (figure 1.1C and D), toxic megacolon and, ultimately, death (Kachrimanidou & Malisiovas, 2011). The current primary method of treatment of CDI is withdrawal of the causative agent if it has been caused by the use of antibiotics, and prescription of other antibiotics, usually metronidazole or vancomycin (Postma *et al.*, 2015).

Separate studies have shown that 30-day mortality rates in the UK are over 30% (McGowan *et al.*, 2011) or even as high as 42% (Wiegand *et al.*, 2012). In Spain, there

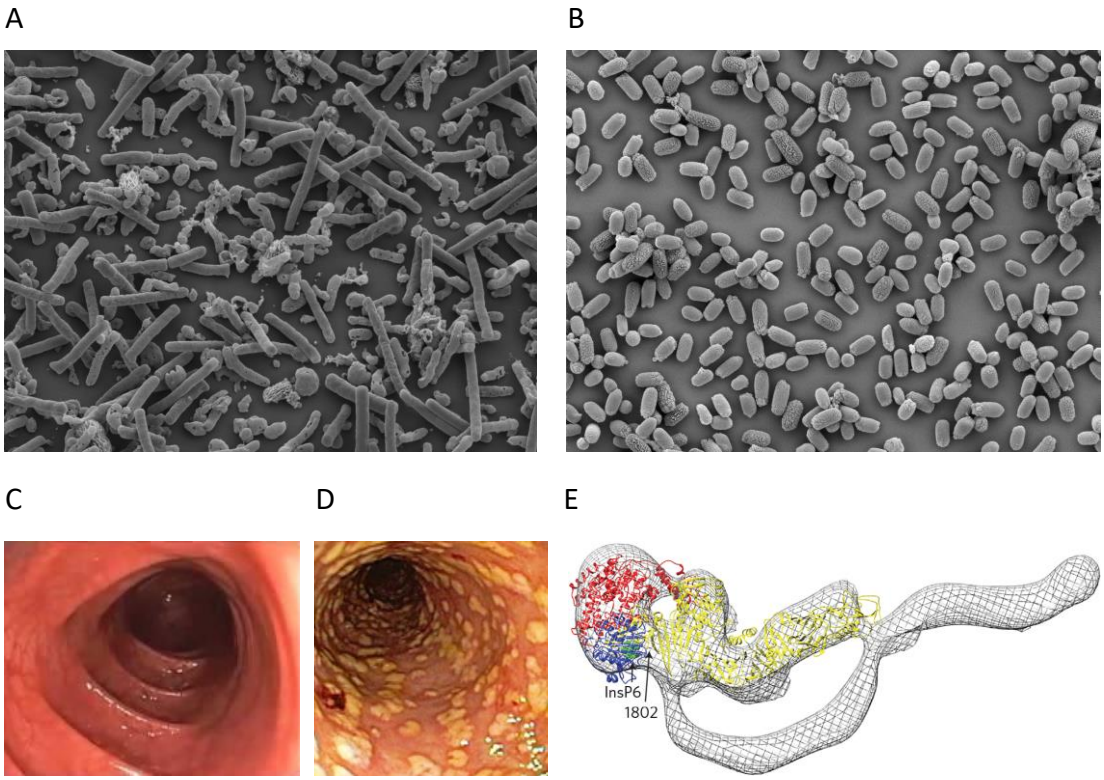


Figure 1.1 Images of *C. difficile.* (A) Electron micrograph showing *C. difficile* Cells. © Public Health England. **(B)** Electron micrograph showing *C. difficile* spores. © Public Health England. **(C)** The appearance of a healthy colon. © American College of Gastroenterology. **(D)** The appearance of a colon showing signs of colitis caused by *C. difficile* (Carrion *et al.*, 2010). **(E)** Composite structure of TcdA. The crystal structure of residues 1-1832 (ribbon) is overlaid on a full length negative stain EM structure. The Glucosyltransferase domain is shown in red, the autoprotease domain in blue, and the delivery domain in yellow. The CROPS domain was not present in this structure but its approximate shape and positioning relative to the rest of the protein can be seen in the EM structure (Chumbler *et al.*, 2016).

are an estimated 7600 cases each year, with each case costing up to €15,000 to treat, resulting in a total annual cost of €32.2 million (Asensio *et al.*, 2013). In the USA, the number of annual number of cases has been estimated at 500,000, resulting in 15-20,000 deaths (Kachrimanidou & Malisiovas, 2011). Treatment is estimated to cost $5000-7000 per case, for a total cost in 2003 of $1.6 billion (Scott, 2009). There has also been a significant global increase in *C. difficile* antibiotic resistance since the early 1990s, which has led to more cases, greater morbidity and mortality and ever increasing costs (Kachrimanidou & Malisiovas, 2011). This presents a clear need for

greater understanding of *C. difficile* for the development of new methods of preventing and treating infections.

Faecal transplants have received considerable attention as a potential new method of treatment of CDI in recent years and appear to have around a 90% cure rate when used to treat recurrent CDI (Borgia *et al.*, 2015; Li *et al.*, 2016). The method has not become routine, however, due to issues with standardisation of administration procedures, storage of samples for later use and a general level of public aversion to the procedure (Borgia *et al.*, 2015; Costello *et al.*, 2015; Li *et al.*, 2016).

The symptoms of CDI are caused by the toxins produced by the bacterium. *C. difficile* produces two large clostridial toxins (LCTs) - TcdA (figure 1.1E), and TcdB, which are also known as Toxins A and B, respectively. The LCTs possess four domains: a glucosyltransferase domain (GTD) an autoprotease domain (APD) a delivery domain and a combined repetitive oligopeptides (CROPS) domain (Chumbler *et al.*, 2016). The first three domains form a relatively small globular region, while the CROPS domain forms a series of β-strands that produce an extended superhelix. The CROPS domain mediates binding of the toxin to host gut epithelial cells, the delivery domain then forms a pore in the cell membrane, through which the GTD is passed which is then cleaved from the rest of the toxin by the APD. The GTD then glucosylates Rho-GTPases, such as RhoA, Rac1, and Cdc42 (Davies *et al.*, 2011).

As well as the LCTs, *C. difficile* also produces a binary toxin, the two components of which are known as CDTa and CDTb. CDTb facilitates the entry of CDTa into epithelial cells. After this, CDTa ADP-ribosylates G-actin, disrupting the equilibrium between F- and G-actin (Voth & Ballard, 2005; Davies *et al.*, 2011).

The action of the LCTs and the binary toxin results in disruption of the actin cytoskeleton of gut epithelial cells, the maintenance of which relies heavily on Rho-GTPases. The damage caused to the cytoskeleton results in cell death and inflammation of the gut, which leads to the symptoms of CDI. Because of this, the LCTs and binary toxin have received considerable attention as potential drug targets, but other possible targets should also be considered if an effective, wide ranging

treatment exploiting multiple aspects of the bacterium is to be developed. To this end, the Surface layer (S-layer) of *C. difficile*, which was first identified in 1984 (Kawata *et al.*, 1984), has received an increasing level of attention over the last 15-20 years.

1.2 Surface Layers

A typical S-layer consists of a single protein arranged in a two dimensional paracrystalline array, forming the outermost surface of the cell (Sara & Sleytr, 2000; Smarda *et al.*, 2002). S-layers have been observed in hundreds of prokaryotic species, including a diverse range of bacteria and virtually all archaea. An S-layer may allow the surface presentation of other cell wall components, but will, by far, form the majority of the externally presented cell surface (Desvaux *et al.*, 2006). The monomers can be arranged in oblique, triangular, square, or hexagonal lattices with one or two, three, four, or six-fold symmetry respectively, although hexagonal lattices appear to be the most common (Sara & Sleytr, 2000). It is believed that all species of a given genus have the same lattice types (Smarda *et al.*, 2002), although structural similarity of all orders can be very low, even between closely related species (Sleytr & Beveridge, 1999; Emerson & Fairweather, 2009).

S-layer proteins usually have a mass between 40 and 200 kDa and are weakly acidic with relatively large proportions of acidic and hydrophobic residues (Smarda *et al.*, 2002). Glycosylation is common, although not universal (Sara & Sleytr, 2000; Schaffer & Messner, 2017). Wide ranging lattice dimensions have been observed, with centre to centre spacing between 30 and 350 Å and heights between 50 and 250 Å. S-layers with higher orders of symmetry tend to have greater dimensions. 20 to 80 Å pores are frequently formed between the subunits and account for between 30 and 70% of the cell surface (Smarda *et al.*, 2002). A range of methods of attachment to the cell are believed to be used by different species: surface-layer homology (SLH) domains have been shown to be present in many species (Engelhardt & Peters, 1998) and secondary cell wall polymers have also been implicated (Ferner-Ortner *et al.*, 2007).

S-layer proteins can account for 15% of the total protein produced by a cell (Sara & Sleytr, 2000), and a generation time of 20 minutes necessitates the translation of around 500 molecules per second (Smarda *et al.*, 2002). It can therefore be inferred from the high metabolic cost of having an S-layer that they must fulfil significant and essential requirements of the cell. Many important S-layer functions have been demonstrated, they include, but are not limited to: archaeal cell shape determination, molecular sieving, the degradation, transport or storage of nutrients or proteins involved in the same, host cell adhesion and/or invasion, immune system evasion, and protection from competing microorganisms (Sara & Sleytr, 2000). For these reasons, the possibility of using various components of the S-layer of *C. difficile* as drug targets has been explored (Kirk *et al.*, 2017), but a more thorough understanding of the S-layer and its components will be needed if this is to be successful.

1.3 *Clostridium difficile* S-layer

As previously mentioned, most S-layers consist of a single protein. The mature S-layer of *C. difficile*, on the other hand, is largely heterodimeric but may contain over 30 other proteins (Sebaihia *et al.*, 2006; Fagan *et al.*, 2011; Monot *et al.*, 2011). The majority of the S-layer is coded for by a single gene, *slpA*, the protein product of which (SlpA), is cleaved after secretion to form low and high molecular weight subunits (LMW SLP and HMW SLP - previously known as P36 and P47, respectively, based on their approximate masses) (Calabi *et al.*, 2001; Karjalainen *et al.*, 2001). HMW SLP is formed of three putative cell wall binding domains (CWBDs, Pfam 04122, CWB2) (Sebaihia *et al.*, 2006; Fagan *et al.*, 2011; Monot *et al.*, 2011; Fagan & Fairweather, 2014), while LMW SLP has a unique fold (Fagan *et al.*, 2009). The two proteins form a heterodimer on the surface of the cell (Fagan *et al.*, 2009) with HMW SLP forming a lower layer and LMW SLP forming an upper, surface exposed layer (Figure 1.2) (Cerquetti *et al.*, 2000). Despite indications to the contrary from early studies (Cerquetti *et al.*, 1992; Mauri *et al.*, 1999; Cerquetti *et al.*, 2000), the S-layer of *C. difficile* does not appear to be glycosylated (Qazi *et al.*, 2009).

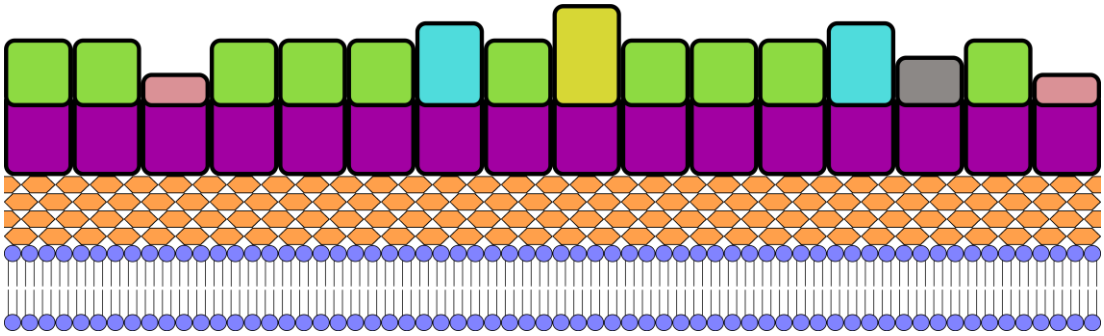


Figure 1.2 Schematic representation of the S-layer of *C. difficile*. The lipid bilayer is shown in blue with the peptidoglycan in peach. Above this is a purple layer formed by the three CWB2 domains of HMW SLP and paralogues. The surface exposed "functional" regions are shown on top, the majority of which are LMW SLP, shown in green. The S-layer also contains other proteins with a range of functions.

The gene *slpA* sits in a 36.6 kb (strain 630) region of the *C. difficile* genome, known as the *slpA* locus. This locus contains 11 *slpA* paralogs (Figure 1.3) and there are 17 more paralogs scattered throughout the genome (Sebaihia *et al.*, 2006; Fagan *et al.*, 2011; Monot *et al.*, 2011). Each of these genes code for a protein with an N-terminal signal peptide and three cell wall binding domains with between 27% and 38% identity to HMW SLP (Calabi *et al.*, 2001; Karjalainen *et al.*, 2001). These paralogs are known as "cell-" or "clostridial wall proteins", or, more commonly, by the abbreviated form "CwpX" (X = 1 - 29). Four *cwp*s (*slpA*, *cwp66*, *cwp84* and *cwpV*) were characterised and named before this convention was established (Fagan *et al.*, 2011). As well as the characteristic three CWB2 domains, most Cwps also possess at least one other domain, allowing the *C. difficile* S-layer to potentially possess an unusually wide range of functions (Figure 1.4). Many of the Cwps are, however, yet to be properly characterised, meaning that an encompassing model of the structure and functions of the S-layer is yet to be established. The intrinsic importance of S-layers combined with their inherent accessibility and the apparent complexity of the S-layer of *C. difficile* may therefore potentially yield a plethora of information that could be exploited for the prevention and treatment of *C. difficile* infections.

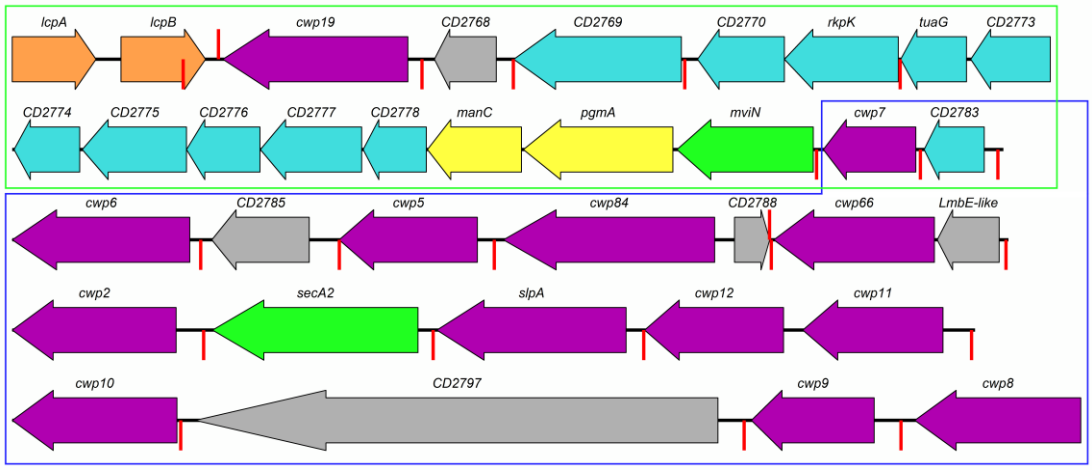


Figure 1.3 The AP and slpA loci from *C. difficile* 630. The two adjacent loci, which respectively code primarily for proteins involved in the production of PSII and proteins that attach to PSII through their CWB2 domains, are shown. The AP locus, as identified by Chu *et al.* (2016), is indicated by a green box while the SlpA locus, as identified by Calabi *et al.* (2001), is indicated by a blue box. *slpA* itself is in the centre of the fourth row. The putative glycosylation cluster identified by Dingle *et al.* (2013), which is not present in strain 630, is found at the end of the third row between *cwp66* and the *LmbE-like* gene, which are switched. When this gene cluster is present, *cwp2* is not. Genes coding for proteins with CWB2 domains are shown in purple, those involved in carbohydrate metabolism in cyan, attachment to peptidoglycan in peach, mannose biosynthesis in yellow and biopolymer export in green, other functions are in grey. CD2768 – hydrolase, CD2785 – membrane protein, CD2788 – GtrA-like membrane protein, CD2797 – calcium binding adhesin. Terminators predicted by Genome2D (Baerends *et al.*, 2004) are shown in red.

Many of the genes within the *slpA* locus show significant variation between strains, particularly in areas that code for the surface exposed regions that vary between paralogues, referred as “functional” regions (Savariau-Lacomme *et al.*, 2003; Fagan *et al.*, 2011; Reynolds *et al.*, 2011; Kirk *et al.*, 2017) as they perform the function of the protein while thy are anchored to the cell wall by the CWB2 domains. There also appears to be a high degree of variability in relative levels of expression between strains (Ferreira *et al.*, 2017). *slpA*, *cwp66*, and *secA2*, which are almost contiguous and appear to usually undergo horizontal transfer as a group, have been noted as having particularly high variation for genes within the *slpA* locus (Dingle *et al.*, 2013): for example, the variable region of *cwp66* has been observed as having as little as 33% identity between strains (Karjalainen *et al.*, 2001). It has been demonstrated that

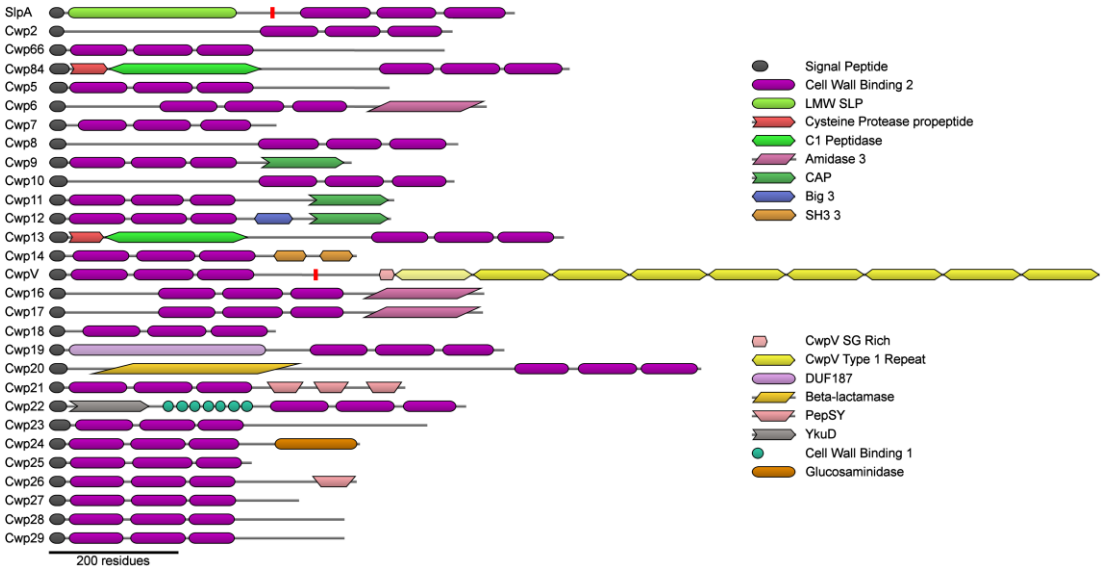


Figure 1.4 Putative domain representation of the 29 cwp genes found in the *C. difficile* 630 genome. Each codes for three cell wall binding domains, while all except cwp18, cwp25, and potentially cwp7 appear to code for at least one other domain, which is likely to confer a specific function on the protein. Many still possess portions for which a function is yet to be determined. SlpA and CwpV cleavage sites are indicated by a red bar. Domains were identified with HMMER and the figure was produced with DoMosaics (Eddy, 2008; Moore *et al.*, 2014).

strain 630 expresses the first seven Cwps at the very least (Calabi *et al.*, 2001) and presents Cwp2, Cwp84, Cwp6, Cwp12, CwpV, Cwp24 and Cwp25 on the cell surface under normal growth conditions (Wright *et al.*, 2005). Interestingly, despite their expression, Cwp66 and Cwp5 were not present in cell surface extracts.

As well as containing the first 12 of the 29 *cwp* genes, the *slpA* locus also contains 6 other genes: 2 putative membrane proteins of unknown function, a putative LmbE-like deacetylase, a non-redundant accessory Sec gene, a putative calcium-binding adhesion protein, and a putative glycosyltransferase (Sebaihia *et al.*, 2006; Monot *et al.*, 2011). The accessory Sec gene - *secA2* - has been demonstrated to be necessary for the secretion of at least some Cwps (Fagan & Fairweather, 2011), although there is a significant possibility that it is required for all of them. It has also been suggested that each of the non-*cwp* genes within the *slpA* locus and several others in the immediately downstream anionic polymer (AP) locus may be involved in cell wall synthesis (Calabi *et al.*, 2001; Willing *et al.*, 2015).

Biazzo *et al.* (2013) analysed 14 of the other 17 *cwp* genes scattered throughout the *C. difficile* genome, amplification of *cwp14*, *cwp21*, and *cwp23* was unsuccessful, so they were excluded from the study. They observed that *cwp13*, *cwpV* (with the exception of the repeat regions, discussed later), *cwp16*, *cwp18*, *cwp19*, *cwp20*, *cwp22*, *cwp24* and *cwp25* have well conserved sequences and expression, suggesting that they may possess important functions. *cwp17*, *cwp26*, *cwp27*, *cwp28*, and *cwp29* tended to be less conserved with considerable variation in expression levels between ribotypes, even when the genes possessed identical sequences (Biazzo *et al.*, 2013). This, along with the fact that *cwp27*, *cwp28*, and *cwp29* are not present in certain ribotypes, suggests that these genes may possess less important functions.

To develop a full model of the workings of the S-layer a thorough understanding of the role of each protein is needed. What follows is a discussion of what is known about each protein and potential roles of their domains. Each protein can be compared to its schematic in figure 1.4.

1.3.1 SlpA

SlpA is the primary component of the *C. difficile* S-layer and is by far the most abundant constituent of cell surface extracts (Wright *et al.*, 2005). It is cleaved after secretion to produce two proteins: HMW SLP and LMW SLP, which form the heterodimeric "H/L complex" (Fagan *et al.*, 2009), this polymerises to form the mature S-layer. HMW SLP binds to the cell wall through a non-covalent interaction (Willing *et al.*, 2015), while LMW SLP is presented as the outermost surface of the cell (Calabi *et al.*, 2001) and appears to have a major role in attachment to host cells (Merrigan *et al.*, 2013). LMW SLP can be extracted from *C. difficile* through relatively gentle methods while the removal of HMW SLP requires harsher conditions (Wright *et al.*, 2005).

The cell wall binding domains of HMW SLP and other Cwps bear low similarity to LytB and LytC, two proteins from *Bacillus subtillis* (Calabi *et al.*, 2001). LytB is an N-acetylmuramic acid L-alanine amidase, also known as a peptidoglycan amidohydrolase, which cleaves peptidoglycan crosslinks between N-acetylmuramic acid and L-alanine. LytC modulates the activity of LytB and may also possess amidase

activity (Lazarevic *et al.*, 1992). HMW SLP exhibits some amidase activity (Calabi *et al.*, 2001), but it is unknown if this function is related to cell wall synthesis or binding. It is also unknown if the CWBDs from other Cwps also possess amidase activity. N-acetylmuramic acid L-alanine amidases have also been shown to bind teichoic acids, polysaccharides embedded in bacterial cell walls (Herbold & Glaser, 1975; Lazarevic *et al.*, 1992).

Using Cwp2 and Cwp66, Willing *et al.* (2015) demonstrated that the three cell wall binding domains present in HMW SLP and all other Cwps mediate attachment to the cell surface through an interaction with PSII, a surface bound polymer formed of a repeating hexasaccharide. They showed that, despite their similarity, the three CWB2 domains are not redundant - each is required for binding to PSII. Removing individual domains, replacing them with a second copy of another or altering their order prevents binding to the cell wall. They also claimed that binding to PSII is mediated through a conserved Pro, Ile/Leu/Val, Ile/Leu/Val, Ile/Leu/Val or "PILL" motif. Although this method of binding is very likely to be used by all Cwps, it has been demonstrated that different methods of S-layer extraction will yield different combinations of Cwps, which suggests there may be slight variations in binding mechanism or strength (Wright *et al.*, 2005).

Despite a high level of variability in the SlpA gene (Dingle *et al.*, 2013), including HMW SLP having a mass between 41 and 48 kDa (Calabi *et al.*, 2001), antibodies raised against HMW SLP from one ribotype retain activity against HMW SLP from another (Cerquetti *et al.*, 2000; Karjalainen *et al.*, 2001). LMW SLP, on the other hand, which can range from 32-38 kDa and has no significant similarity to any other proteins (Calabi *et al.*, 2001), is not always recognised by antibodies raised against another ribotype. This variability is likely to have arisen in an attempt to evade the host immune system (Cerquetti *et al.*, 2000; Calabi *et al.*, 2001; Spigaglia *et al.*, 2011), which is also likely to be the reason why other Cwps show an increased level of variability between strains (Dingle *et al.*, 2013). Variations in LMW SLP have also been shown to be likely to result in changes in the ability of *C. difficile* to adhere to host cells (Merrigan *et al.*, 2013).

The first insights into the structure of the S-layer of *C. difficile* were obtained by Cerquetti *et al.* (2000) (figure 1.5A and B) who used two different methods to visualise the S-layer by scanning electron microscopy. This demonstrated that two separate layers are formed, a lower one with apparent hexagonal symmetry formed by HMW SLP and an upper one with apparent square symmetry formed by LMW SLP (Cerquetti *et al.*, 2000). These images, however, did not yield much structural detail beyond determination of the symmetry of the S-layer. Fagan *et al.* (2009) analysed the structure of the H/L complex using small angle X-ray scattering (SAXS) and determined the crystal structure of a fragment of LMW SLP at 2.4 Å. This structure

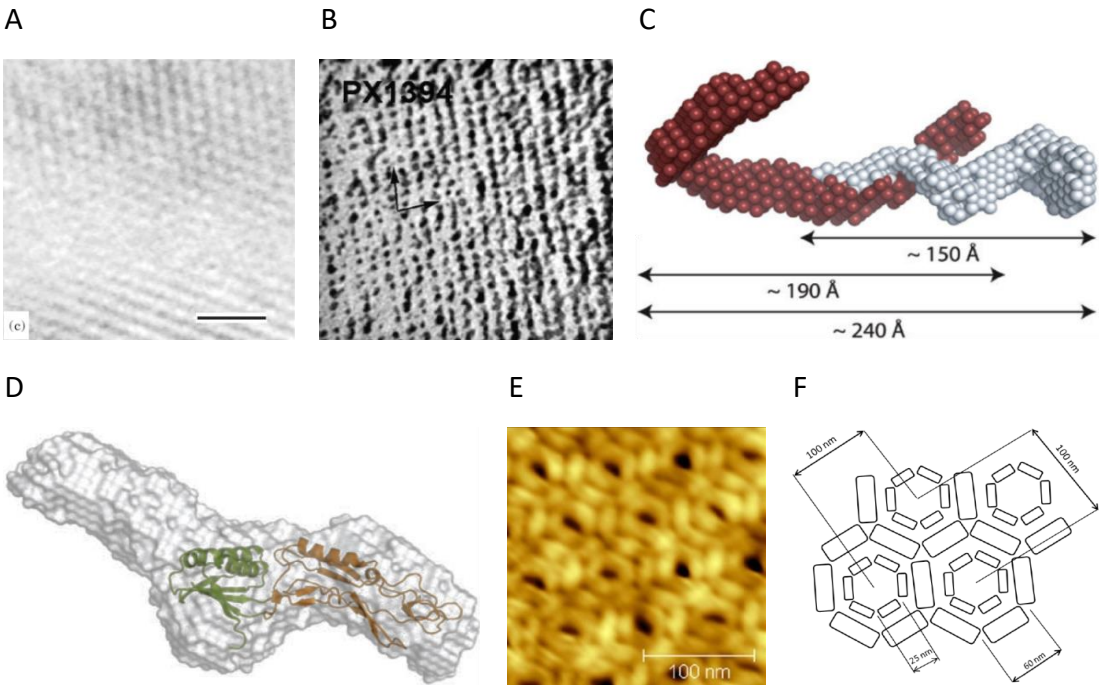


Figure 1.5 Current structural insights into SlpA and the S-layer as a whole. (A) Negatively stained TEM of reassembled non-purified urea extracted S-layer proteins showing hexagonal symmetry. **(B)** Freeze etched cell wall extract showing square symmetry (Cerquetti *et al.*, 2000). **(C)** SAXS structure of the H/L complex, HMW SLP is shown in crimson, LMW SLP is shown in grey. **(D)** Crystal structure of truncated LMW SLP superimposed on the SAXS structure of full length LMW SLP, domain 1 is shown in green and domain 2 in orange, the protein only crystallised when the C-terminal portion, which is responsible for dimerization with HMW SLP, was not present (Fagan *et al.*, 2009). **(E)** AFM topograph of urea extracted S-layer proteins. **(F)** Approximate dimensions of features seen in E (Kirby, 2011).

was of a truncated form missing 59 residues from the C-terminus that are likely to be involved in formation of the H/L complex. The N and C-terminal regions form a small two-layer sandwich, while the central region assumes a novel fold consisting of around 60% loops. As this domain is likely to be surface exposed, it was suggested that the loops allowed for the high level of sequence variability that confers immune system evasion, while retaining the overall fold (Fagan *et al.*, 2009).

The S-layer has also been examined using atomic force microscopy (AFM) (Kirby, 2011). Urea S-layer extracts were reconstituted on mica and a hexagonal array was visualised at near-nanometre resolution (figure 1.5E and F). This confirmed the hexagonal symmetry previously seen by Cerquetti *et al.* (2000) using the same extraction method (figure 1.5A) and significantly improved upon the resolution, revealing a pore with a maximum diameter of approximately 60 nm with spacing between pores of approximately 100 nm. It remains unclear how two separate layers are able to form with differing symmetry while their respective primary components form a heterodimer.

1.3.2 Cwp66

Due to a low level of similarity to known bacterial adhesins, Waligora *et al.* (2001) predicted that Cwp66 is also an adhesin and analysed its ability to perform this function. They observed that Cwp66 is secreted under normal growth conditions and that surface presentation is increased in response to heat-shock. Adherence to Vero cells by heat-shocked *C. difficile* was partially abrogated by antibodies raised against Cwp66, particularly those raised against the likely surface exposed C-terminal functional region, but it was not affected without prior heat-shocking (Waligora *et al.*, 2001).

cwp66 is located 32 bp (strain 630) downstream of the putative LmbE-like deacetylase gene also found in the *slpA* locus (Figure 1.3). The two genes have no separating terminator or promoter, so are polycistronically co-transcribed. The LmbE-like superfamily consists of a wide range of metallohydrolases, the majority of which bind zinc as a cofactor. All members of the family possess a Rossmann fold and cleave substrates containing an N-acetylglucosamine moiety. Many LmbE-like

proteins have been shown to possess cell wall related functions, so the family is of particular interest for drug development (Viars *et al.*, 2014). It has previously been suggested that Cwp66, the LmbE-like deacetylase and the adjacent Cwp2 may assemble to form an adhesive complex on the surface of the cell (Savariau-Lacomme *et al.*, 2003), however, this hypothesised complex is yet to be observed.

A PDB BLAST shows that the functional domain of Cwp66 bears no significant similarity to any previously determined folds (Altschul *et al.*, 1990). It contains three imperfect 21-23 residue repeats and is predicted to assume a structure mostly comprised of β-strands (Waligora *et al.*, 2001).

1.3.3 Cwp84 and Cwp13

Cwp84 and Cwp13 each possess a C1A cysteine protease domain (also known as a papain protease domain). Cwp84 is responsible for the cleavage of SlpA to form HMW SLP and LMW SLP (Karjalainen *et al.*, 2001; Kirby *et al.*, 2009; Dang *et al.*, 2010). It has also been shown to be capable of breaking down gelatine and the extra cellular matrix proteins fibronectin, laminin, and vitronectin, but is unable to cleave type IV collagen (Janoir *et al.*, 2004; Janoir *et al.*, 2007). Cwp84 knockouts present full length SlpA on the surface of the cell. This results in an abnormal S-layer (Kirby, 2011) and the presence of SlpA, Cwp2 and Cwp66 in growth media, which is not seen in the wild type. Knockouts also show aberrant colony morphology, grow at half their usual rate, and have a propensity to aggregate (Kirby *et al.*, 2009; de la Riva *et al.*, 2011). A Cwp84 knockout strain was, however, still able to cause CDI in hamsters (Kirby *et al.*, 2009), but it has been suggested that perturbation of S-layer formation may make the bacterium more susceptible to antibiotics (Dang *et al.*, 2010).

Despite a high level of identity to Cwp84, Cwp13 appears to possess different functions and is not as essential to correct functioning of the cell (de la Riva *et al.*, 2011). While Cwp84 cleaves SlpA between LMW SLP and HMW SLP, Cwp13 cleaves it within one of the cell wall binding domains, rendering the protein useless. It has been speculated that this function may facilitate the removal of misfolded protein, ensuring a fully functional S-layer (de la Riva *et al.*, 2011).

Papain proteases possess an N-terminal propeptide and are frequently, but not always, able to autoactivate (Nagler *et al.*, 1999; Dahl *et al.*, 2001; ChapetónMontes *et al.*, 2011; Beton *et al.*, 2012). de la Riva *et al.* (2011) showed that Cwp84 is unlikely to be capable of autoactivation, while Cwp13 is likely to be able to autoactivate, the structural basis for this difference is currently unclear. Cwp13 was also shown to be capable of removing the propeptide from Cwp84, although it does not appear to be entirely responsible for this as Cwp13 knockouts present both the proenzyme and mature Cwp84 on the surface of the cell.

1.3.4 Cwp6, Cwp16 and Cwp17

Unlike the rest of the family, which possess either N- or C-terminal cell wall binding domains, those of Cwp6, Cwp16 and Cwp17 are central within the protein rather than at either of the termini. The three proteins have been predicted to possess an amidase 3 domain at the C-terminus, while the predictions performed using HMMER and Interpro (Eddy, 2008; Jones *et al.*, 2014) were unable to assign a structure for a region of approximately 150 residues at the N-terminus (figure 1.4). The effect that the positioning of CWB2 domains, whether N-terminal, C-terminal, or indeed, central, has on the overall structure of Cwps, and their positioning relative to the cell wall and therefore their interactions with PSII is unknown.

The recently determined structure of Cwp6 confirmed the predicted C-terminal amidase domain and also showed the presence of a seven-stranded β-barrel at the N-terminus, which is also likely to be present in Cwp16 and Cwp17 (Usenik *et al.*, 2017) (figure 1.6). The β-barrel bears a high level of structural similarity to the runt homology domain from the RUNX family of eukaryotic transcription factors. The RUNX family of proteins are a group of metazoan transcription factors whose functions can be modulated via a wide range of posttranslational modifications and have been shown to be frequently downregulated in cancer (Ito *et al.*, 2015). Heterodimeric RUNX proteins appear to act as weak transcriptional repressors on their own, but when complexed with other proteins can act as considerably stronger activators or repressors (Durst & Hiebert, 2004). It does not appear that prokaryotic runt domains have been previously observed, so the role of this domain in Cwp6,

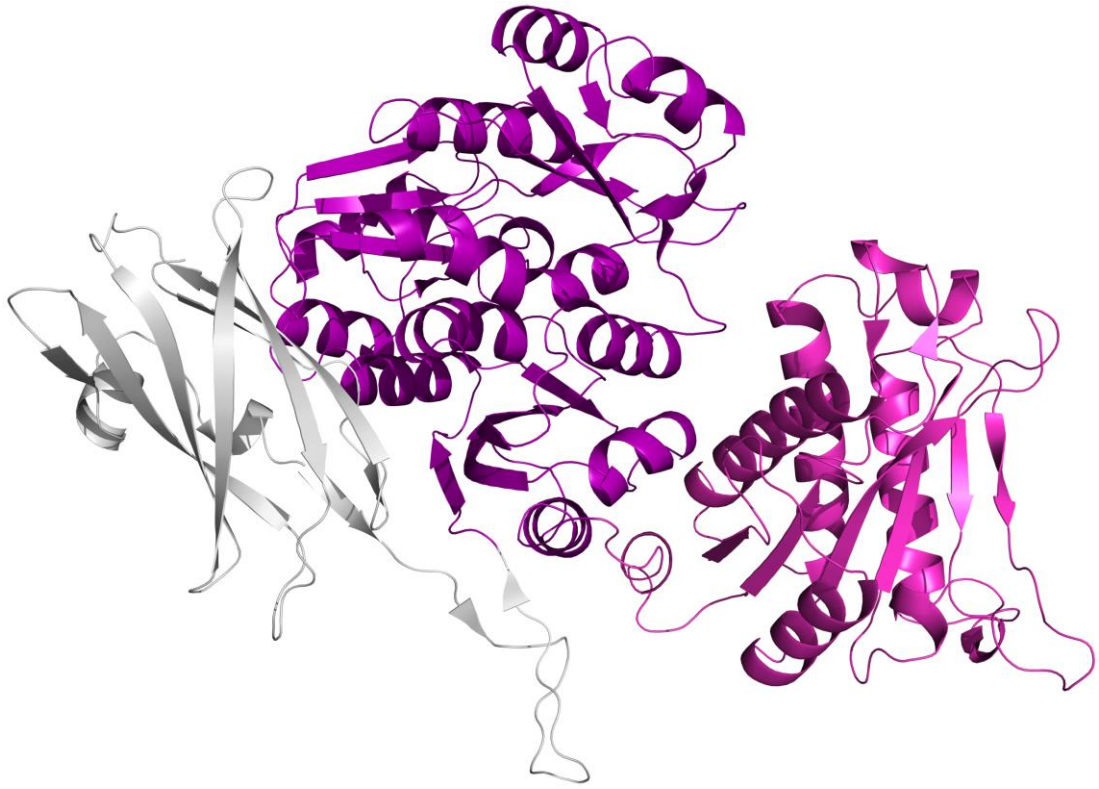


Figure 1.6 The structure of Cwp6. (Usenik *et al.*, 2017) The N-terminal Runt domain is coloured grey, the central CWB2 domains are purple and the C-terminal amidase domain is lilac. Central CWB2 domains are only seen in Cwp6 and the closely related Cwp16 and Cwp17. The effect that this has on PSII binding and therefore orientation on the surface of the cell is yet to be determined. The function of the usually metazoan runt domain is also currently unknown.

Cwp16 and Cwp17 is unclear. As the eukaryotic domains are involved in a significant number of protein-protein interactions, this may be the case in prokaryotes too.

Amidase 3 domains possess N-acetylmuramic acid L-alanine amidase activity – that is to say they are capable of cleaving the bond between N-acetylmuramic acid and L-alanine in peptidoglycan crosslinks (Senzani *et al.*, 2017). The knockout of an amidase 3 containing protein from *Mycobacterium smegmatis* recently showed impaired cell division, increased susceptibility to antibiotics and increased cell permeability (Senzani *et al.*, 2017). An ability to break down peptidoglycan was demonstrated for Cwp6 (Usenik *et al.*, 2017), however, as previously noted, HMW SLP has also been shown to possess amidase activity (Calabi *et al.*, 2001). Whether the amidase activity

shown by Cwp6 was conferred by the amidase 3 domain, the cell wall binding domains, or both, was not considered.

1.3.5 Cwp9, Cwp11 and Cwp12

The N-terminal cell wall binding domains of Cwp12 are followed by a type 3 bacterial immunoglobulin-like domain (Big 3) and a CAP domain (Eddy, 2008) (named after the related mammalian Cysteine-Rich Secretory Proteins, insect Antigen 5 proteins, and plant Pathogenesis-Related proteins) (Gibbs *et al.*, 2008). Despite bearing 63% identity and 80% similarity to Cwp12 (Altschul *et al.*, 1990), a Big 3 domain is not detected in Cwp11 by an HMM search (Eddy, 2008). Based on the high degree of similarity to Cwp12, it is likely that it does possess a Big 3 domain but it is simply not detected due to the low sequence similarity frequently seen in Big 3 domains (Bateman *et al.*, 1996). Cwp9 is approximately 75 residues shorter as it does not contain a Big 3 domain.

Big domains, which are likely to have evolved either divergently or horizontally from eukaryotic immunoglobulins (Bateman *et al.*, 1996), are frequently found on the surface of bacterial cells (Wang *et al.*, 2013) and have been shown to be involved in host cell adhesion and invasion (Hamburger *et al.*, 1999; Luo *et al.*, 2000; Czibener & Ugalde, 2012). Despite particularly low sequence similarity, all members of the family are predicted to have largely similar structures (Bateman *et al.*, 1996; Wang *et al.*, 2013).

The first structure of a Big 3 domain, that of *Streptococcus pneumoniae* SP0498, was published in 2013 (Wang *et al.*, 2013). Big 3 domains consist of an eight stranded stretched β-barrel, a somewhat different structure to that of eukaryotic immunoglobulins, which possess more of a β-sandwich (Wang *et al.*, 2013). SP0498 was demonstrated to be a calcium binding protein, a feature that is potentially common to all Big domains (Raman *et al.*, 2010). It was speculated that calcium binding is important to the role of Big domains in host cell adhesion and invasion (Wang *et al.*, 2013).

In eukaryotes, CAP domains are involved in a wide range of signalling processes and their roles have been extensively studied. Members of the superfamily have an α-β-α sandwich fold and appear to function through a conserved "incomplete protease" active site containing two histidine residues and an acidic residue (usually glutamate) (Gibbs *et al.*, 2008). The wide range of functions exhibited by CAP domains is usually conferred through another domain or a C-terminal extension (Brangulis *et al.*, 2015).

Despite being widespread, prokaryotic CAP domains are yet to be as extensively characterised as their eukaryotic homologues. Brangulis *et al.* (2015) determined the structure of BB0689, a surface presented CAP domain from *Borrelia burgdorferi* that has a potential role in pathogenesis, and performed a range of assays to identify the function of the protein. The study showed that BB0689 possesses the conserved features seen in eukaryotic CAPs and therefore, potentially many bacterial CAPs do, but the authors were unable to identify any function.

1.3.6 Cwp14

Cwp14 contains N-terminal cell wall binding domains and two domains that are classified by Pfam as bacterial SH3 domains, which are also known as type 3 SH3 domains (Finn *et al.*, 2016), while InterPro classifies the domains as SH3-like domains (Jones *et al.*, 2014).

SH3 (Src Homology 3) domains, named after the Rous Sarcoma Virus tyrosine kinase, v-Src (Thomas & Brugge, 1997), to which they have significant sequence similarity, are 50-60 residue domains that form a 5 or 6 stranded beta sandwich with a hydrophobic ligand binding pocket capable of binding proteins with a PXXP motif. The domain facilitates a wide range of protein-protein interactions across all organisms and has a very large range of functions (Weng *et al.*, 1995; Mayer, 2001).

1.3.7 CwpV

CwpV has N-terminal CWB2 domains followed by a region of approximately 200 residues of unknown structure and function, a short Ser/Gly rich region and several repeat regions. The sequence of CwpV is very well conserved between ribotypes up to and including the Ser/Gly rich region (Reynolds *et al.*, 2011). The protein has been

shown to mediate cell aggregation and phage resistance. Overexpression results in smaller, more densely packed colonies and decreased susceptibility to infection by bacteriophages, while knocking down or knocking out results in larger, sparser colonies and increased susceptibility (Reynolds *et al.*, 2011; Sekulovic *et al.*, 2015). The mechanisms by which CwpV causes aggregation and phage resistance are currently unknown, however, two particularly interesting features have been observed: firstly, the level of expression is controlled by phase variability of the gene (Emerson *et al.*, 2009), secondly, the repeat regions are highly variable but appear to retain their function (Reynolds *et al.*, 2011).

CwpV is expressed by 0.1-10% of *C. difficile* cells, regardless of descent from a common parent cell and accounts for approximately 13% of the S-layer (Reynolds *et al.*, 2011). Expression is controlled by the recombinase RecV, which inverts a pair of imperfect inverted repeat regions located between the promoter and the start codon. This results in two possible mRNA transcripts, one that results in translation of CwpV (termed “ON”), and one that does not (“OFF”). The differences between the two transcripts result in the formation of a stable stem loop intrinsic terminator structure in the OFF transcript that is not formed in the ON transcript. When RNA polymerase reaches the intrinsic terminator, transcription is stopped and the complex destabilised, preventing transcription (Emerson *et al.*, 2009).

Five completely unrelated repeat types of approximately 80-120 residues have been identified in various ribotypes. CwpV is able to mediate aggregation and phage resistance regardless of which repeat regions it contains. Strains have been observed with between 4 and 9 repeat regions, accounting for roughly 50-75% of the residues within the protein. The five types of repeats bear no significant similarity to each other but each show a high degree of similarity between multiple copies within a protein. The first copy of a repeat is generally afforded slightly greater sequence variability (Reynolds *et al.*, 2011). Two CwpV type II repeats are also seen in the sortase substrate CD3392 (Peltier *et al.*, 2017).

It has previously been observed that CwpV may undergo some form of cleavage, however it was unclear how this cleavage was mediated (de la Riva *et al.*, 2011).

Dembek *et al.* (2012) determined that CwpV autoproteolyses into two fragments via N-O acyl migration. The cleavage site, Gly412-Thr413, is roughly half way between the CWB2 domains and the Ser/Gly rich region. Asp411 deprotonates Thr413, which then nucleophilically attacks Gly412, forming a hydroxyoxazoladine intermediate (Dembek *et al.*, 2012). This is reduced to an ester, and then hydrolysed to produce the cleaved products: an N-terminal product of approximately 42 kDa, and a C-terminal product of up to 120 kDa (Reynolds *et al.*, 2011). The extreme chemical conditions normally required for N-O acyl rearrangements are believed to be made unnecessary by unusual torsion of Asp411. The mechanism was confirmed by a series of mutations (Dembek *et al.*, 2012). The two products have been shown to co-elute, so it is likely that they form a non-covalent complex, with the highly conserved regions either side of the cleavage site potentially forming the interface between the two cleavage products (Reynolds *et al.*, 2011). It is currently unknown if there is any similarity between this interface and the one within the H/L complex.

1.3.8 Cwp19

Cwp19 is predicted to possess an N-terminal glycoside hydrolase-like 10 (GHL 10) domain, separated from the C-terminal CWBDs by approximately 70 residues (Altschul *et al.*, 1990; Finn *et al.*, 2016). The function of Cwp19 is currently unknown, however attempts have been made to determine the structure of the protein (Kirby *et al.*, 2011).

The gene coding for Cwp19 is located in the anionic polymer locus (AP locus) (figure 1.3) which is likely to be involved in synthesis of PSII, the polysaccharide that mediates binding of CWB2 domains (Sebaihia *et al.*, 2006; Monot *et al.*, 2011; Willing *et al.*, 2015; Chu *et al.*, 2016). It is therefore possible that Cwp19 may be involved in processing PSII.

1.3.9 Cwp20

Cwp20 possesses an N-terminal region of unknown structure and function of approximately 60 residues followed by a β-lactamase domain, another region of unknown structure and function of around 320 residues and C-terminal cell wall binding domains.

β-lactamases are the most widely studied group of antibiotic resistance enzymes. They were discovered in 1940, before β-lactam antibiotics (including penicillins, cephalosporins, monobactams, carbapenems and others) entered clinical use (Abraham & Chain, 1940). They now serve as the primary antibiotic resistance mechanism in gram-negative bacteria. β-lactamases are a diverse group of antibiotic resistance enzymes; many species express several, resulting in resistance to a wide range of β-lactam antibiotics (Liakopoulos *et al.*, 2016). There are currently 17 known β-lactamases or penicillin binding proteins in the *C. difficile* genome, including Cwp20, which makes the therapeutic use of β-lactams difficult (Sebaihia *et al.*, 2006; Monot *et al.*, 2011).

1.3.10 Cwp21 and Cwp26

Cwp21 features N-terminal CWB2 domains followed by three PepSY domains while Cwp26 is predicted to contain one C-terminal PepSY domain separated from the CWB2 domains by an uncharacterised region of approximately 120 residues (Eddy, 2008). PepSY domains, which derive their name from peptidase and *Bacillus subtilis* YpeB, are usually 60-75 residues long, are believed to act as protease inhibitors and are frequently (though not always) found in protease propeptides. Sequence conservation among PepSY domains is usually very low with only a central aromatic residue and an aspartate flanked by two hydrophobic residues with a nearby glycine residue showing a high level of conservation, although even these are not always present. It has been speculated that secreted proteins containing PepSY domains may play a role in controlling the bacterium's environment and pathogenesis (Yeats *et al.*, 2004).

1.3.11 Cwp22

Cwp22 contains a YkuD domain followed by 8 type 1 cell wall binding (CWB1) repeats (Eddy, 2008). YkuD domains, which were previously known as ErfK/YbiS/YcfS/YnhG domains, are now named after a protein from *B. subtilis*, the first in the family to have its structure determined (Bielnicki *et al.*, 2006). YkuD domains are L,D-transpeptidases, which appear to perform roles similar to the more common D,D-transpeptidase involved in peptidoglycan crosslinking. The reversal of

stereochemistry seen in L,D-transpeptidases is believed to confer resistance to β-lactam antibiotics (Biarrotte-Sorin *et al.*, 2006). The proteins are composed of a β-sandwich and possess a conserved active site consisting of a (Y/L)XXHG(S/T) motif closely followed by SXGC(I/V)R(M/L), with the histidine, first glycine, cysteine and arginine forming a catalytic tetrad.

The 20 residue CWB1 repeats, which have been seen in a wide range of proteins from gram-positive bacteria, assume a β-hairpin fold and contain conserved hydrophobic residues, aromatic residues and glycines (Fernandez-Tornero *et al.*, 2001). Successive β-hairpins are orientated at approximately 120° to each other, resulting in a left-handed superhelix. CWB1 repeats are found in choline binding proteins and glucosyltransferases (Shah *et al.*, 2004). Both the choline and the carbohydrate binding sites are formed by the interface between adjacent hairpins (Fernandez-Tornero *et al.*, 2001). Interestingly, these repeats are similar to those found in the CROPS domain of the large clostridial toxins (Davies *et al.*, 2011).

1.3.12 Cwp24

Cwp24 has N-terminal CWB2 domains followed by a region of unknown structure and function of approximately 60 residues and a C-terminal Glycoside hydrolase family 73 domain, specifically, an endo-β-N-acetylglucosaminidase domain. This is predicted to cleave between N-acetylglucosamine (NAG) and N-acetylmuramic acid (NAM) in peptidoglycan (Eddy, 2008; Jones *et al.*, 2014; Finn *et al.*, 2016). This could either be for remodelling of the *C. difficile* cell wall, or for attacking competing bacteria.

1.3.13 Uncharacterised regions

Despite the wide range of putative domains currently identified, twelve Cwps, namely, Cwp2, 66, 84, 5, 8, 13, 20, 23, 26, 27, 28, and 29 each contain regions of around 100 residues or more for which no structure or function has so far been predicted. This leaves a large number of potential functions of the S-layer still to be determined.

1.3.14 SecA2

The secretory pathway is responsible for the majority of protein translocation across gram-positive cell walls. Proteins possessing a signal peptide are passed through the SecYEG channel by the ATPase activity of SecA, frequently after recognition by the signal recognition particle (SRP), a ribonucleoprotein complex (Driessen & Nouwen, 2008; du Plessis *et al.*, 2011; Zhou *et al.*, 2014). It had previously been believed that bacteria possessed only one copy of each of the *sec* genes, however, in recent years, an increasing number of species have been shown to possess a second copy of *secA*, *secY*, or both. These genes are referred to as accessory *sec* genes (Rigel & Braunstein, 2008; Feltcher & Braunstein, 2012). They are usually not essential to the survival of the bacterium and are only responsible for a small portion of the secretosome – frequently proteins involved in pathogenicity.

A study by Fagan & Fairweather (2011) characterised *C. difficile*'s accessory *secA* gene. This gene is found in the *slpA* locus and, by convention, is known as *secA2*. It was demonstrated that neither of the two SecAs is redundant and that SecA2 is necessary for the secretion of at least SlpA, Cwp2, Cwp66, Cwp84 and CwpV. *secA2* knockouts, which would presumably be unable to form an S-layer, were not viable as the S-layer is likely to be essential to viability. *secA2* knockdowns, which were shown to have compromised SlpA and CwpV secretion, were viable but severely stressed (Fagan & Fairweather, 2011). This strongly indicates that the signal peptides of at least the identified Cwps, if not all, are sufficiently different to a typical signal peptide that they are unable to bind to SecA. The exact method by which SecA and the SRP recognise proteins for secretion has only recently begun to be elucidated (Grady *et al.*, 2012; Zhou *et al.*, 2014).

1.4 Summary

Clostridium difficile kills tens of thousands of people each year and antibiotic resistance is increasing. New strategies for prevention and treatment of *C. difficile* infections (CDI) and new drug targets are needed to combat this. The S-layer of *C. difficile* is one such potential target (Kirk *et al.*, 2017).

S-layers possess two properties that make them ideal drug targets: they are surface exposed and they are essential to survival. The S-layer of *C. difficile* is likely to possess a larger number of proteins compared to other S-layers, allowing it to perform a wide range of functions. A detailed understanding of the S-layer of *C. difficile*, its various components, and their respective structures and functions will be required if it is to be exploited for prevention and treatment of CDI. The aim of this work was to elucidate the structures of some of the proteins in the S-layer, Specifically Cwp84, Cwp2 and Cwp19, primarily using X-ray crystallography, and to use these structures to better understand the functions of the respective proteins, thereby adding to understanding of the S-layer as a whole.

Chapter 2

X-ray Crystallography

2.1 Structural Biology

It has been argued that we are unable to fully understand physiological processes without knowledge of the molecular structures involved, and even less able to manipulate them (Harrison, 2004). By studying the structure of a protein we are able to make inferences about its function and how it is able to perform said function. When combined with other techniques, including biochemical, biophysical, bioinformatic and cellular, structural biology can be a very powerful tool for understanding how biological systems function and, potentially, for determining ways that that function can be artificially modulated (Yee *et al.*, 2005; Zheng *et al.*, 2014). Proteins generally have a size in the order of tens of nanometres, while visible light has a wavelength in the order of hundreds of nanometres. Because of this, despite recent technology pushing the boundaries, the structural information that can be gleaned by studying proteins using visible light based techniques, i.e. light microscopy, is severely limited (Heintzmann & Ficz, 2013). Several techniques do exist, however, for studying the structure of objects at scales smaller than those permitted by visible light. The three techniques that are used to determine the structure of proteins at an atomic or near-atomic scale are X-ray crystallography, nuclear magnetic resonance (NMR) and cryo electron microscopy (cryo-EM) (Harrison, 2004).

Crystallography is frequently referred to as the “gold standard” for structural biology (Zheng *et al.*, 2014; Bond, 2015). Molecules of any size can be studied using the technique as long as the molecule of interest can be crystallised and data of sufficient quality can be collected. In practice, this usually limits the technique to molecules with a maximum mass in the order of hundreds of kilodaltons (kDa), while NMR is limited to molecules of a size of up to around 40 kDa, but these limits are constantly being pushed (Yee *et al.*, 2005; Reddy *et al.*, 2010). NMR also does not involve the often time-consuming repeated attempts to crystallise the molecule being studied, which are sometimes never successful and yields better information about the state of the molecule in solution (Yee *et al.*, 2005). Cryo-EM, on the other hand, is more limited by how small the molecule is. It is usually used to study complexes with sizes

ranging from hundreds of kilodaltons to several megadaltons, however, recent improvements in techniques have reduced the minimum size of molecules that can be visualised by cryo-EM and the resolution that can be achieved is also constantly improving (Scheres, 2014; Cheng, 2015).

Generally speaking, crystallography is usually able to determine structures to a higher resolution than NMR or Cryo-EM. 68% of all crystal structures deposited in the Protein Data Bank (PDB) have resolutions between 1.5 and 2.5 Å, with 98% being between 1.0 and 3.5 Å (Berman *et al.*, 2000). Structures at resolutions beyond approximately 1.0 Å are considered to be of atomic resolution and even at this resolution yield limited information on hydrogen atoms present in the sample, due to the small number of electrons. For this, diffraction experiments should be performed using neutrons, rather than X-rays. Neutron diffraction is rarely able to achieve the resolution that is possible using X-rays but interactions are independent of atomic number so hydrogen is visible, even at low resolution (Blakeley *et al.*, 2015).

Small angle X-ray scattering (SAXS) and SANS, the analogous technique using neutrons, allow a range of statistics relating to the behaviour of the molecule of interest in solution to be determined and it is also possible to use the statistics generated to calculate the probable shape of the molecule to a low resolution. SAXS can also be a powerful tool for examining conformational differences, especially when combined with higher resolution techniques (Schnablegger & Singh, 2013).

The work described in this thesis uses X-ray crystallography as a primary technique. What follows is a brief summary of the theory behind the technique.

2.2 What is a Crystal?

The determination of a structure using X-ray crystallography, by definition, requires that the substance of interest be crystallised before said determination can be performed. “A crystal is an anisotropic, homogenous body consisting of a three-dimensional periodic ordering of atoms, ions or molecules” (Borchardt-Ott, 2011). The nature of this periodicity, however, can vary significantly between crystals, which

affects how they interact with X-rays. To be able to determine the structure, we must understand the arrangement of the molecules within each crystal, which results in the diffraction pattern.

The simplest unit within a crystal, from which the whole of the crystal can be generated by a series of defined symmetry operations, is known as the asymmetric unit. The space group of a crystal defines what symmetry operators can be applied to the asymmetric unit to generate the unit cell. The unit cell is a parallelepiped that repeats by simple translations to generate the crystal. All copies of the asymmetric unit within a crystal must be in identical environments. For this to occur, copies of the unit cell must be capable of tessellation. Considering two dimensional shapes, it becomes clear that the only orders of symmetry within the unit cell that are permitted are none, two-fold, three-fold, four-fold, and six-fold, as these are the only forms of symmetry that generate shapes capable of tessellation. Excluding mirror symmetry, which is not permitted for chiral molecules such as proteins, this equates to six relatively trivial two-dimensional plane groups, generated by rotations around a central point of 360°/N, with N being the order of symmetry (Rupp, 2010).

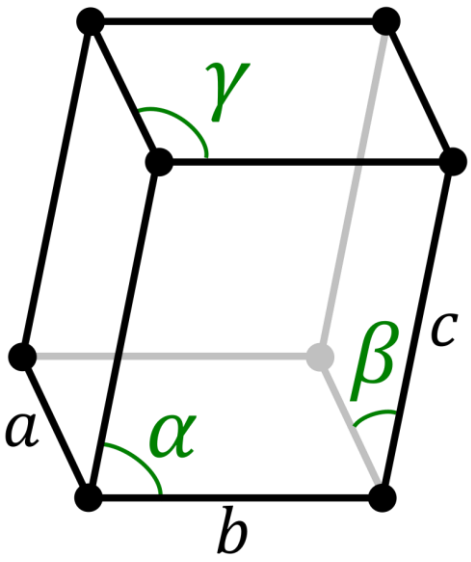


Figure 2.1 The unit cell. Regardless of symmetry, all unit cells will be parallelepipeds. At higher degrees of symmetry, more restrictions are placed upon the vectors that make up the dimensions and angles of the unit cell.

In three dimensions, the symmetry is still restricted by these six tessellating two dimensional shapes. The shapes can now be described using three dimensions: a, b and c and three angles: α, β and γ (figure 2.1). As every copy of the asymmetric unit must be in an identical environment and because certain symmetry operations around one point will result in symmetry around another, there are certain restrictions on the values of a, b, c, α, β and γ dependent on the symmetry within the crystal. This results in seven possible crystal systems with certain restrictions on the properties of the lattice. Crystals that belong to each of the crystal systems may still have differing forms of symmetry though, so they are then divided into 230 three-dimensional space groups, however, as proteins are chiral, they cannot assume space groups with mirror symmetry, glide planes or inversions, so only 65 space groups apply (Rupp, 2010). These space groups are listed in table 2.2.

2.3 Solving a Structure

2.3.1 Crystallisation

It goes without saying that to determine the crystal structure of a protein, the protein must first be crystallised. By far the most common method currently used for crystallisation is vapour diffusion, which is widely used to screen hundreds of crystallisation conditions. A vapour diffusion experiment can either take the form of a hanging drop or a sitting drop experiment. In either, a small sample of protein is mixed with a solution that frequently contains a salt, a buffer and a precipitant, but may contain many other compounds. This drop is placed in a sealed chamber with a reservoir of a significantly larger volume of the solution being screened. As the reservoir will contain a much higher concentration of precipitant (double, if protein and reservoir solution are mixed in a 1:1 ratio), net vapour diffusion will drive water from the drop to the reservoir (figure 2.3). The result of this is a gradual increase of protein and precipitant concentration. This pushes the protein to a supersaturated state, at which point it will either crystallise or precipitate (Rupp, 2010).

Crystal system	**Lattice properties**	**Space groups**	**z**
Triclinic	$a \neq b \neq c$ $\alpha \neq \beta \neq \gamma \neq 90°$	P1	1
Monoclinic	$a \neq b \neq c$ $\alpha = \gamma = 90°$ $\beta \neq 90°$	$P2,\ P2_1$	2
		C2	2
Orthorhombic	$a \neq b \neq c$ $\alpha = \beta = \gamma = 90°$	$P222,\ P222_1,\ P2_12_12,\ P2_12_12_1$	4
		$I222,\ I2_12_12_1$	4
		$C222,\ C2_12_12_1$	4
		F222	4
Tetragonal	$a = b \neq c$ $\alpha = \beta = \gamma = 90°$	$P4,\ P4_1,\ P4_2,\ P4_3$	4
		$I4,\ I4_1$	4
		$P422,\ P42_12,\ P4_122,\ P4_12_12,\ P4_222,$ $P4_22_12,\ P4_322,\ P4_32_12$	8
		$I422,\ I4_122$	8
Trigonal	$a = b \neq c$ $\alpha = \beta = 90°$ $\gamma = 120°$	$P3,\ P3_1,\ P3_2$	3
		R3	3
		$P312,\ P321,\ P3_112,\ P3_121,\ P3_212,\ P3_221$	6
		R32	6
Hexagonal		$P6,\ P6_1,\ P6_2,\ P6_3,\ P6_4,\ P6_5$	6
		$P622,\ P6_122,\ P6_222,\ P6_322,\ P6_422,$ $P6_522$	12
Cubic	$a = b = c$ $\alpha = \beta = \gamma = 90°$	$P23,\ F23,\ I23$	12
		$P2_13,\ I2_13$	12
		$P432,\ P4_232$	24
		$F432,\ F4_132$	24
		I432	24
		$P4_332,\ P4_132,\ I4_132$	24

Table 2.2 (previous page) List of space groups. The 65 chiral space groups are shown sorted into their crystal systems with associated lattice properties. "z" is the number of copies of the asymmetric unit within the unit cell for the given space groups (Rupp, 2010).

The crystallisation of a substance is driven by the change in entropy, ΔS. The second law of thermodynamics states that in an isolated system, entropy increases, $\Delta S > 0$. A single crystallisation experiment is not an isolated system, or even a completely closed system, however the second law can still be held to be true. The entropy of a substance in solution will always be greater than the entropy of the crystallised substance as a crystal is ordered, while a solution is disordered, therefore, when a protein crystallises, $\Delta S_{protein} < 0$. This means that the change in entropy of the protein will never drive crystallisation. Instead, we must consider the change in entropy of the solution that the protein is in. While the protein is in solution, each molecule will be surrounded by an ordered, low entropy hydration shell. As the

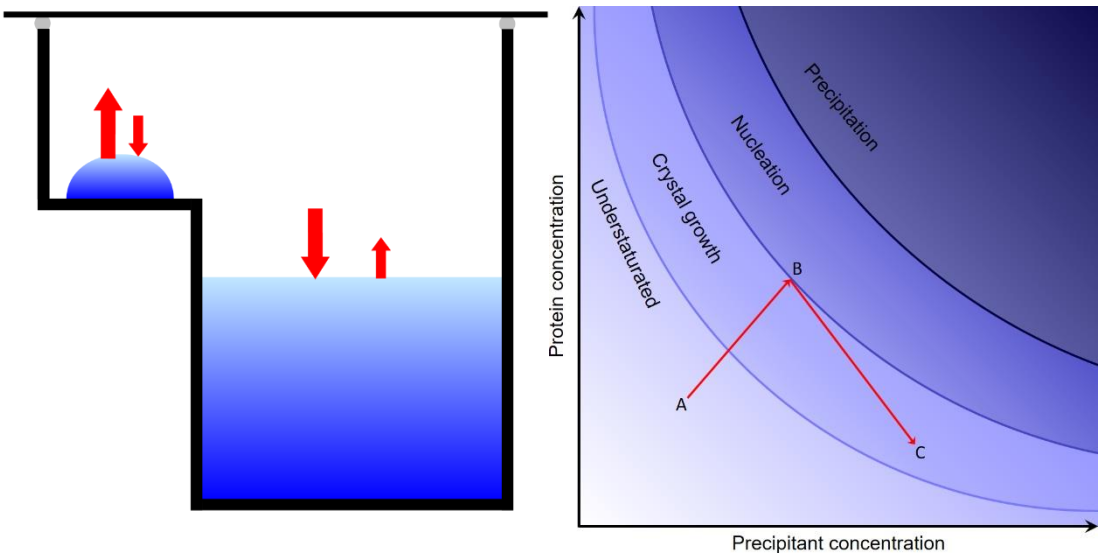


Figure 2.3 The process of crystallisation. A schematic of a sitting drop experiment is given on the left. The drop contains a mixture of protein solution and the crystallisation condition being screened. A much larger volume of this condition is contained within the reservoir. As the reservoir contains a higher concentration of precipitant, net diffusion will occur from the drop to the reservoir, raising the protein concentration within the drop. The diagram on the right shows what happens as this occurs. Starting at A, the protein is at an under saturated concentration. As the concentration increases, it enters the nucleation zone, at which point, if entropy allows, crystals begin to nucleate (B). During this time, the protein is still over saturated so crystals will grow until an equilibrium is reached (C).

protein crystallises, the majority of the molecules in hydration shells will be released, allowing their entropy to increase, $\Delta S_{solution} > 0$. From this it is evident that ΔS will only be greater than zero if $\Delta S_{solution} > \left|\Delta S_{protein}\right|$. For the vast majority of conditions, this is not the case. For this reason, we must screen a wide number of conditions to find ones that our protein will crystallise in (Derewenda & Vekilov, 2006).

2.3.2 X-ray Data Collection

X-ray diffraction data can be collected from a crystal either at a synchrotron or using a home source. Home sources tend to generate X-rays by bombarding a rotating anode with electrons. This results in the loss of inner shell electrons, which are replaced by higher energy outer shell electrons with the concomitant release of X-rays at wavelengths characteristic of the anode material (Rupp, 2010). Synchrotrons accelerate electrons in a many-sided polygon. At each vertex, the beam is bent, the angular deceleration of the electron beam results in the production of bremsstrahlung (German for roughly "breaking radiation") with an intensity orders of magnitude higher than that produced by home sources. Even higher intensity radiation can be produced on the straight sections through the use of an undulator, a device containing a series of magnets of alternating polarity which produces waves within the electron beam, again producing bremsstrahlung (Motz, 1951). Once the X-ray beam has been focused and, in the case of a synchrotron, a specific wavelength selected, the crystal, mounted on a goniometer, is exposed to the beam. The goniometer allows precise positioning of the crystal within the beam and rotation of the crystal during data collection (Rupp, 2010). The generation of a diffraction pattern (figure 2.4) relies on constructive interference between parallel planes within a crystal (figure 2.5). This constructive interference only occurs and therefore spots only appear on the diffraction pattern if the diffraction satisfies Bragg's law (Bragg, 1913):

$$n\lambda = 2d \sin\theta$$

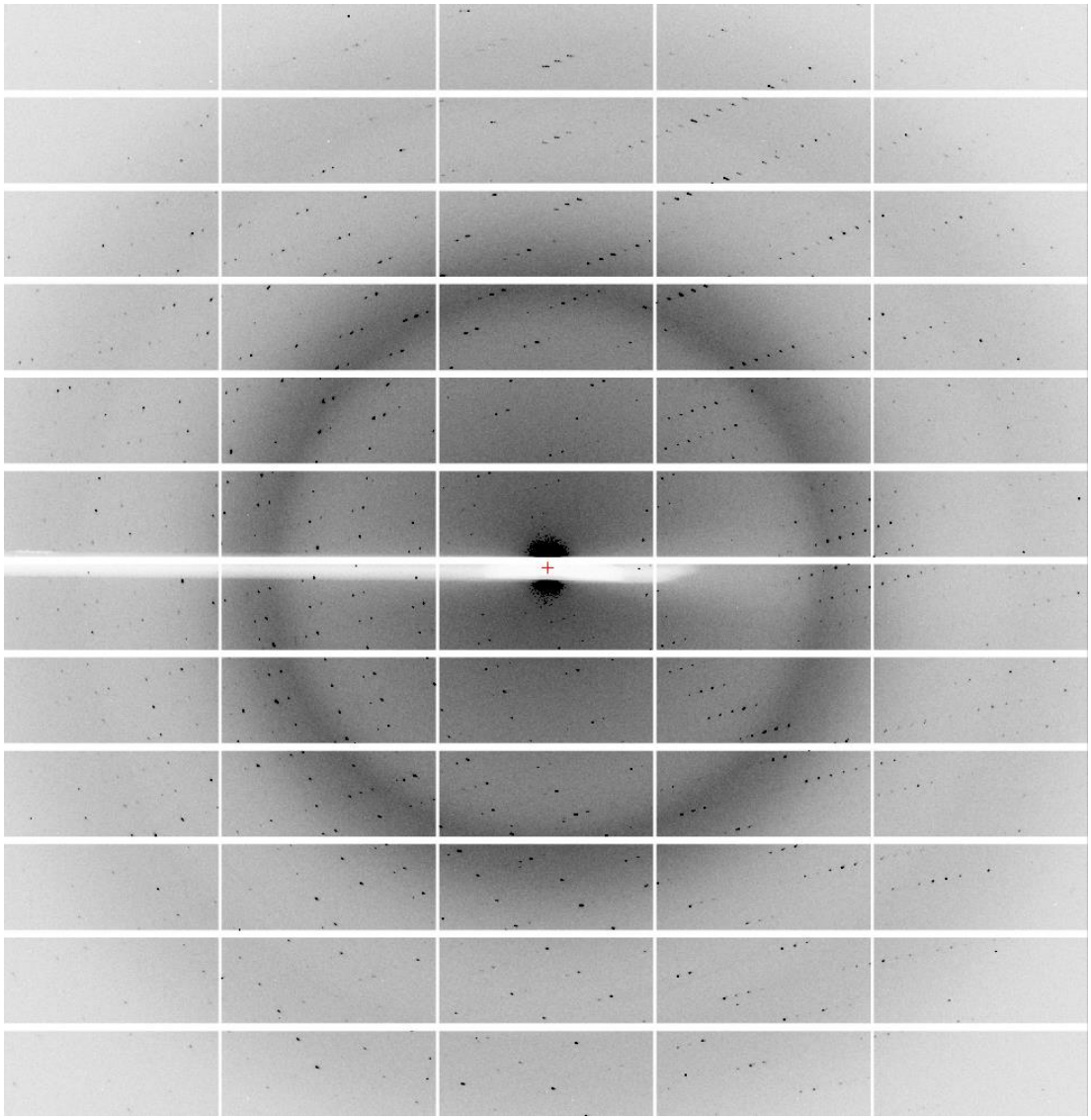


Figure 2.4 Example diffraction pattern. Each spot results from diffraction from a set of planes within the crystal that satisfy Bragg's law. Spots further from the centre of the image give higher resolution data. The diffraction pattern is a 1.7 Å test image of human angiogenin mutant R95Q collected at Diamond Light Source (Bradshaw *et al.*, 2017b).

That is to say, constructive interference occurs when the sine of the angle of incidence on the diffracting planes times double the distance between the two planes is equal to a multiple of the wavelength of the incident X-ray beam (figure 2.5). It can be noted from this that $\sin\theta$ is inversely proportional to the plane separation. Therefore, a crystal with a large unit cell will produce more reflections that satisfy Bragg's law with smaller angles of incidence. This may initially seem counter intuitive,

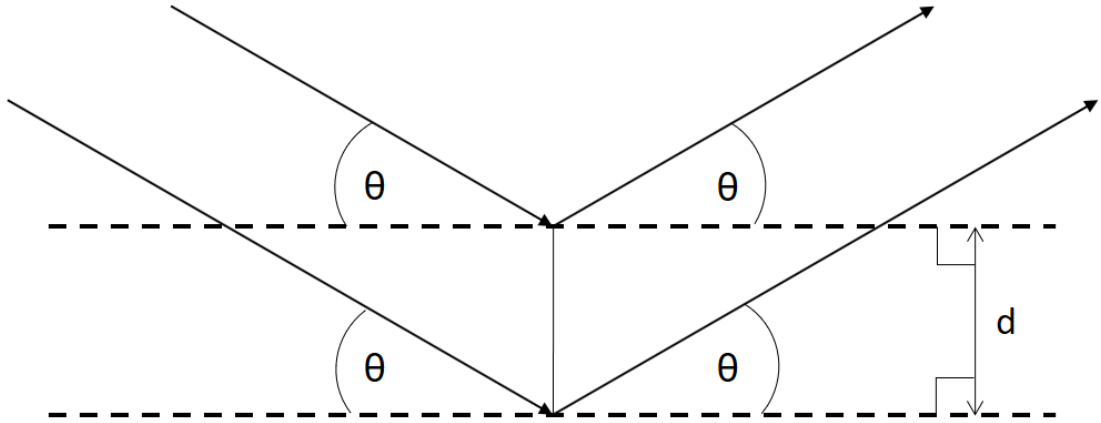


Figure 2.5 Diffraction satisfying Bragg's law. Constructive interference will occur and a spot will be produced in a diffraction pattern if the diffraction of the X-ray beam of wavelength "λ" incident at angle "θ" diffracted by multiple planes separated by distance "d" satisfies Bragg's law. In reality, the value of θ is much smaller than the angle shown here.

however, it is logical that, at a given resolution, a crystal with greater cell dimensions will contain more information per unit cell (Rhodes, 1993).

Planes within crystals were known about long before the discovery of X-rays (Miller, 1839; Rontgen, 1896). Planes can be described according to their Miller index (*hkl*). Each set of planes that satisfies Bragg's law will produce a single reflection. However, certain symmetry elements within the crystal will result in destructive interference at specific Miller indices, preventing a spot from appearing. These "missing spots" are known as "systematic absences" and their presence, or lack thereof, can aid in indexing a diffraction pattern. If symmetry elements that produce systematic absences are present, the Miller indices of all observed reflections must satisfy the conditions given in table 2.6 (Rhodes, 1993; Rupp, 2010).

After diffraction has occurred, diffracted X-rays then hit a detector. Early X-ray detectors used film, this has since been successively replaced by image plates (IPs), charged-coupled devices (CCDs) and finally by pixel-array detectors (PADs). Detector technology is constantly improving, with modern PADs such as PILATUS or EIGER, produced by DECTRIS, capable of collecting tens or hundreds of images per second (Somogyi *et al.*, 2015).

Table 2.6 Reflection conditions. When certain symmetry elements are present, systematic absences will occur in the diffraction pattern. All observed reflections will then satisfy the conditions given within the table (Rupp, 2010).

Symmetry element	Reflection condition
2_1 along a, b, or c	*h00*, *0k0* and *00l* must be even
4_2, or 6_3, along c	*00l* must be even
3_1, 3_2, 6_2, 6_4 along c	*00l* must be a multiple of 3
4_1, 4_3, along a or c	*h00* and *00l* must be multiples of 4
6_1 or 6_5, along c	*00l* must be a multiple of 6
C	*h* + *k* must be even
I	*h* + *k* + *l* must be even
F	*h*, *k* and *l* must be all even or all odd

The crystal is rotated during the diffraction experiment in an attempt to observe as many of the reflections as possible, that is to say, to ensure that diffraction occurs off nearly all unique planes within the crystal.

2.3.3 Data Processing

2.3.3.1 Indexing

All diffraction can be assigned to a triclinic space group, in the case of a chiral substance, such as a protein, this can only mean P1. However, it is likely that higher orders of symmetry will be present. 22.5% of all structures in the PDB are in the most common space group $P2_12_12_1$, while only 4.9% are in P1, the sixth most common space group (Berman *et al.*, 2000).

Once diffraction data have been collected they are assigned to P1 and initial cell dimensions are calculated from the images using stronger reflections. This can usually be done with a single image but larger numbers of images are frequently used to resolve ambiguities and determine somewhat more precise cell dimensions, particularly if there are a small number of reflections on each image (Rupp, 2010). Higher symmetry space groups are then considered. Each solution is scored for how well the data appear to fit it based upon the lattice properties given in table 2.2 and

the reflection conditions given in table 2.6. The user then selects what they believe to be the most likely solution based upon this score and how common each space group is (Rupp, 2010). P1 will be given the lowest score, meaning that the data fit well into the space group. Higher symmetry space groups will be given gradually higher scores as more restrictions are placed upon the data and slight errors within the data prevent these restrictions from being exactly met. Eventually, the score will jump significantly when the data completely fail to meet the restrictions of a particular space group. Generally, the scores for each indexing solution are examined and the solution with the highest symmetry before the jump in score is selected. Ultimately, however, the only way to determine if this solution is correct is to solve the structure of the protein (Rupp, 2010).

Once a space group has been selected and cell dimensions determined, the location of all reflections on the images can be predicted. Modern data collection strategies using pixel array detectors employ a fine-slicing technique that result every reflection being spread across multiple images (Mueller *et al.*, 2012). Each observed reflection has its intensity recorded and catalogued according to its Miller index. Historically, this was a very "tedious" process (Perutz, 1949) but it is now considerably easier thanks to automation through programs such as MOSFLM (Abrahams & Leslie, 1996; Battye *et al.*, 2011), *XDS* (Kabsch, 2010) and *DIALS* (Waterman *et al.*, 2016).

2.3.3.2 Scaling and Merging

Once the data have been indexed and integrated, they are scaled. This is necessary as, despite attempts to ensure uniformity of experimental parameters throughout the experiment, many of them may vary very slightly between images. Examples of parameters that need to be corrected for include exposure time, beam geometry, crystal to detector distance, crystal orientation and absorption of x-rays by the crystal. Scaling compares images and adjusts intensities accordingly (Evans, 2006). Reflections observed across multiple images are also combined into a single measurement. The data are then merged, that is to say, multiple observations of reflection with a given set of Miller indices have their intensities averaged in an attempt to determine the "true" intensity of each reflection. The quality of the scaled

and merged data must then be assessed. This process is frequently iterative, with scaling statistics resulting in the removal or addition of weak or poorer data by a combination of image number and resolution.

Various programs are available to make scaling a trivial process, notably *AIMLESS* (Evans & Murshudov, 2013) and its predecessor, *SCALA* (Evans, 2006), however, assessment of the quality of a dataset and determination of its nominal resolution, are not. A number of statistics are presented by scaling programs to this end.

2.3.3.3 Data Quality

A series of statistics known as "R-factors" are probably the most widely used determinants. The first of these was R_{sym} (Arndt *et al.*, 1968) which is closely related to the more commonly used R_{merge}:

$$R_{merge} = \frac{\sum_{hkl} \sum_{i=1}^{n} |I_i\,(hkl) - \bar{I}(hkl)|}{\sum_{hkl} \sum_{i=1}^{n} I_i\,(hkl)}$$

R_{merge} gives a measure of the spread of the intensities of a series of separate measurements of a reflection around the average of those measurements, summed for all reflections within a dataset (Karplus & Diederichs, 2012): the lower the value of R_{merge}, the more consistent, and therefore, the better the data. R_{merge} was, arguably, the most important descriptor of data quality for decades, although more recently there has been a significant level of argument against the use of R_{merge} to indicate data quality (Diederichs & Karplus, 1997; Weiss & Hilgenfeld, 1997; Evans & Murshudov, 2013; Karplus & Diederichs, 2015). It was noted that as the multiplicity of a dataset, i.e. the number of times the average reflection is observed, increases, R_{merge} also increases. Logically, the greater the number of independent measurements of a reflection that are used to determine the intensity of that reflection, the closer the average measured intensity will be to the "true" intensity. Therefore, higher multiplicities are better, however, the resulting higher value of R_{merge}, which tends to infinity with multiplicity, regardless of data quality (Evans &

Murshudov, 2013) implies poorer data (Weiss & Hilgenfeld, 1997). To attempt to remove this contradiction, R_{meas} was devised (Diederichs & Karplus, 1997):

$$R_{meas} = R_{rim} = \frac{\sum_{hkl} \sqrt{\frac{n}{n-1}} \sum_{i=1}^{n} |I_i\,(hkl) - \bar{I}(hkl)|}{\sum_{hkl} \sum_{i=1}^{n} I_i\,(hkl)}$$

R_{meas} is larger than R_{merge} by an approximate factor of the root of the multiplicity divided by the multiplicity minus one, ultimately becoming equal to R_{merge} for a dataset with infinite multiplicity. This prevents R_{meas} from becoming unduly large for high multiplicity datasets, relative to those of lower multiplicity. R_{meas} does, however, still increase with multiplicity, tending to infinity. To combat this, R_{pim} was introduced (Weiss, 2001):

$$R_{pim} = \frac{\sum_{hkl} \sqrt{\frac{1}{n-1}} \sum_{i=1}^{n} |I_i\,(hkl) - \bar{I}(hkl)|}{\sum_{hkl} \sum_{i=1}^{n} I_i\,(hkl)}$$

For a dataset with a particularly low multiplicity, R_{pim} is approximately equal to R_{merge}, however for higher multiplicities, and therefore theoretically more accurate datasets, R_{pim} decreases, but this decrease is offset by high error, allowing a relatively accurate determination of a high resolution cut off. Despite the ability of these statistics to determine the overall quality of a dataset, it is still claimed that they are not particularly good at determining the quality of weak data (i.e. data in the outer shells, at the highest resolution) and therefore, can make the determination of a suitable high resolution cut-off, the resolution beyond which the errors in data make the data useless, difficult (Karplus & Diederichs, 2015). To this end, the Pearson's product-moment correlation coefficient between two randomly split halves of a dataset - $CC_{1/2}$ - can be used (Assmann *et al.*, 2016).

$$CC_{1/2} = \frac{\sum(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\left(\sum(a_i - \bar{a})^2 \sum(b_i - \bar{b})^2\right)}}$$

"Where a_i and b_i are the intensities of unique reflections merged arcoss the observations randomly assigned to subsets A and B, respectively, and $\bar{a}$ and $\bar{b}$ are their averages". It has been demonstrated that data may become no better than random at a $CC_{1/2}$ of 0.27 (Evans & Murshudov, 2013). Accordingly, a $CC_{1/2}$ of 0.5 to 0.3 in the outer shell is commonly used as a high resolution cut-off, however it has also been argued that a $CC_{1/2}$ as low as 0.1 may still result in the inclusion of useful data (Karplus & Diederichs, 2012; Diederichs & Karplus, 2013; Karplus & Diederichs, 2015). Conservative cut-offs will result in the loss of data and, at the very least, including weak data does not appear to negatively impact upon the final structure (Evans & Murshudov, 2013).

Another statistic frequently used to determine data quality and an approximate resolution limit is the signal to noise ratio:

$$\langle |I|/\sigma(I) \rangle = \frac{1}{N} \sum \frac{|I_i|}{\sigma(I_i)}$$

That is, the average intensity of all observations of a unique reflection, divided by their standard deviation, summed for a set of reflections, divided by the number of reflections (Rupp, 2010). This value is lower for weak data and it has been claimed that data with a signal to noise ratio of 0.9 are no better than random (Evans & Murshudov, 2013).

The completeness of data should be taken into consideration as well. With a sub-optimal data collection protocol, useful reflections may not be recorded. Having

shells with particularly low completeness can lead to failure of phasing or streaking in the electron density map, which can ultimately lead to problems in refinement. There is some disagreement over an acceptable minimum level of completeness, however it has been stated that completeness below 70-80% can be severely deleterious (Rupp, 2010).

The statistics described here are not the only statistics that can be used to determine data quality. They do, however, appear to be some of the most widely used and ultimately, a good understanding of what each statistic tells us combined with knowledge of what is required of the data will allow careful modifications to scaling and merging protocols so that the highest quality information can be gleaned from a given dataset. Once this has been achieved, the data can be put into the electron density equation (Rupp, 2010):

$$\rho(x,y,z) = \frac{1}{V}\sum_{hkl} |F_{hkl}| e^{2\pi i \varphi_{hkl}} e^{-2\pi i(hx+ky+lz)}$$

This equation calculates the electron density that resulted in the observed diffraction, which is a measure of the likelihood of the presence of electrons at a given set of coordinates: $\rho(x,y,z)$. This is determined by the inverse Fourier transform: $\sum_{hkl} e^{-2\pi i}$ at given reciprocal space plane coordinates: $(hx+ky+lz)$, of the structure factors for each reflection with Miller index hkl: $|F_{hkl}|e^{2\pi i \varphi_{hkl}}$. With modern computing power, this would be trivial, were it not for the phase problem, discussed in section 2.3.5.

2.3.4 Matthews Coefficient

Before the structure can be solved, we need to know how many molecules of our protein are present in the asymmetric unit. For a protein crystal to form, individual protein molecules must be packed against other protein molecules and largely unable to move. This results in a theoretical minimum protein content within the unit cell

and corresponding maximum solvent content. Conversely, even with very closely packed protein molecules there will always be gaps between the molecules that are filled with solvent, resulting in a theoretical maximum protein content and corresponding minimum solvent content. Because of this, a reliable way to determine the number of protein molecules in the asymmetric unit is to divide its volume by the mass of its protein contents. The resulting number is known as the Matthews coefficient (Rupp, 2010). Matthews (1968) calculated this value for 116 structures and determined that it lies between 1.5 and 6 $Å^3Da^{-1}$ with the corresponding solvent contents lying between 27 and 78%. By performing this calculation for an unknown structure with a series of potential numbers of molecules and comparing the results to known distributions, the possible numbers of molecules, their corresponding solvent contents and, most importantly, the probability that each number is correct can be determined. These probability calculations were updated by Kantardjieff & Rupp (2003) to also factor in the resolution that a crystal has diffracted to. It was shown that crystals with lower solvent contents tend to diffract to higher resolutions as a greater portion of the crystal is ordered. Once the number of molecules within the asymmetric unit is known, attempts can be made to solve the phase problem.

2.3.5 The Phase Problem

Structure factors have an amplitude component, $|F_{hkl}|$, and a phase component, $e^{2\pi i\varphi_{hkl}}$. The amplitude is proportional to the square root of the intensity of each reflection, so is known, while the phase is lost in the experimental process, so must be approximated by other means (Taylor, 2003).

2.3.5.1 Patterson Methods

The Patterson function takes the electron density equation and sets the phase of every reflection to zero by squaring the structure factor (Patterson, 1935). This results in the phase component of every structure factor having a value of one. The Patterson function is therefore, effectively the electron density equation with the phase component removed:

$$P(u,v,w) = \frac{1}{V}\sum_{hkl} |F_{hkl}|^2 e^{-2\pi i(hu+kv+lw)}$$

This produces a three dimensional representation of interatomic distance vectors within the unit cell. As a given vector $\overrightarrow{AB}$ will also have the opposite vector $\overrightarrow{BA}$, the Patterson "map" is centrosymmetric. The interatomic distance vectors can therefore simply be measured from the map and atomic positions determined from a series of simultaneous equations. However, for a unit cell containing N atoms, the corresponding Patterson map will contain N^2 peaks. This places a limit on the number of atoms within the unit cell for which Patterson methods are a viable option for solving the phase problem at around 10 atoms for manual calculation or 100 for computer calculation, which makes this method unsuitable for solving macromolecular structures. The Patterson function can, however, still be used to predict symmetry operators within a crystal and determine the locations of heavy atoms (Rupp, 2010).

2.3.5.2 Direct Methods

Direct methods rely on the prior knowledge that the correct phases will cause the resulting electron density map to not contain any negative density and that all density will be roughly spherical and evenly distributed within the unit cell. This allows relationships to be established between reflections only through which these criteria can be met. Once a small number of phases have been accurately estimated, the remaining phases can be rapidly calculated using these relationships. The atomicity of the density does, however, require data of very high resolution (<1.2 Å) from crystals with very low solvent content, features rarely seen in protein crystals. Direct methods are, therefore, generally only suitable for complete structure determination in small molecule crystallography or for very good data, the likes of which are generally only achievable with particularly small proteins. As with Patterson methods, direct methods can be and are used for heavy atom substructure determination (Taylor, 2003).

2.3.5.3 Molecular Replacement

In the early 1960s, Michael Rossmann realised that “many of the larger protein molecules are made up of identical, or closely similar sub-units”, a fact that could potentially be exploited in solving the phase problem so that heavy atom derivatives would not have to be used. This idea initially applied to molecules related by non-crystallographic symmetry within the same structure, but was later exploited for cases “where an unknown structure is to be solved with a known search molecule” (Rossmann & Blow, 1962; Rossmann, 2001). If the known structure is close enough to the unknown structure, approximate phases can be calculated, first by determining the correct orientation of the molecule through a rotation function, then by determining the positioning of the molecule through a translation function (Rossmann, 1990).

Despite Patterson functions not being suitable for direct calculation of interatomic vectors in macromolecular structures, they can be used to check the validity of a solution. The interatomic vectors given on a Patterson map can be divided into two categories: intramolecular and intermolecular. Intermolecular vectors generally have a greater magnitude than intramolecular vectors, so the two are largely separated on a Patterson map. Intramolecular vectors depend only on the orientation of the molecule, not its positioning, which allows a series of possible orientations of the molecule to be sampled. Hypothetical Patterson maps are calculated and compared to the actual Patterson map for the data, the one that resembles it most closely is assumed to have the most accurate orientation (Rupp, 2010).

Once the respective orientations of all molecules have been determined, possible interatomic vectors can be calculated based on hypothetical positions within the asymmetric unit. Again, a series of translations are sampled and the corresponding Patterson maps are compared to the actual Patterson map to determine which positioning is closest to that in the structure to be determined. Knowing the orientation and positioning of all molecules, hypothetical phases can be calculated.

Patterson functions are still used in molecular replacement by some programs, including MolRep (Vagin & Teplyakov, 1997) and AMoRe (Navaza, 1994), however

Phaser (McCoy *et al.*, 2007), which has seen significant increases in popularity in recent years uses maximum likelihood methods.

Maximum likelihood methods calculate the structure factors (F_{calc}) for a given model (that is, orientation and position of the similar structure being used) and compares them to the structure factors observed in the data (F_{obs}) (Rupp, 2010). The probability that the present model is correct is then calculated. As the model will always have errors, this probability is very low, so is expressed as log-likelihood. The gain in log-likelihood (LLG) between models is used to drive the programs towards the more likely solution. As with Patterson methods, rotation function calculations are performed before translation functions (McCoy, 2007; McCoy *et al.*, 2007).

2.3.5.4 Experimental Methods

The ability of an atom to cause diffraction (known as the atomic scattering factor) varies with the wavelength of the incident radiation, with characteristic wavelengths for each element, known as absorption edges, resulting in significant changes to their scattering factors (Taylor, 2003). This effect is greater for heavier atoms. As proteins are primarily made of hydrogen, carbon, nitrogen and oxygen this effect is very weak in native protein. By incorporating heavier atoms into a structure, we are able to increase this effect. Experimental methods aim to calculate the phases by exploiting differences in the scattering factors caused by the presence of a heavy atom (Taylor, 2003; Rupp, 2010).

The atomic scattering factor can be determined as follows (Blow, 2002; Rupp, 2010):

$$f_\lambda = f_0 - \delta f'_\lambda + i f''_\lambda$$

Where f_0 is the real wavelength independent conventional atomic scattering factor, $\delta f'_\lambda$ is the real change to the scattering factor at the given wavelength, caused by dispersive effects and $i f''_\lambda$ is the imaginary component that results in a 90° change in phase, caused by anomalous effects.

A reflection can be considered to be a vector with a magnitude proportional to the intensity and a direction equal to the phase. As we do not know the phase, this can be graphically represented as a circle of radius equal to the contribution of the protein to the structure factor (figure 2.7):

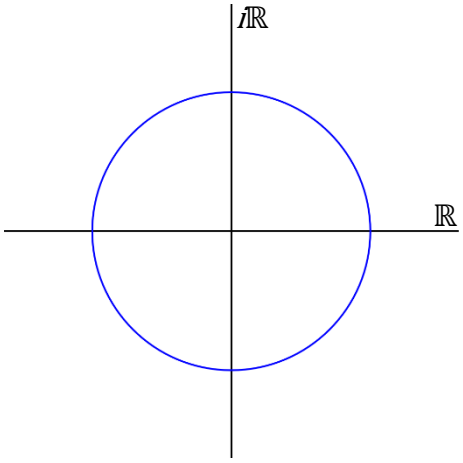


Figure 2.7 Phase contribution of native protein. The intensity of our reflection is known and can be represented as a distance from the origin while the phase, represented as a direction, is not known, so we draw a circle.

In isomorphous replacement (single or multiple – SIR, MIR), which was first used by Perutz (1956) and Kendrew *et al.* (1958), derived from techniques developed by Beevers & Lipson (1934), crystals are soaked in solutions containing heavy atoms. As the location of spots is simply a function of the symmetry of the crystal and the dimensions of the unit cell and independent of its contents, the locations of the spots in a diffraction pattern from a heavy atom derivative should be unchanged compared to the native protein. The intensities, on the other hand, will differ (Taylor, 2003). Simply by subtracting the intensities of the native protein, from those soaked in the heavy atom, we can determine the approximate phase and intensity contributions of the heavy atoms: $|F_H| \approx |F_{PH}| - |F_P|$. If we add a circle of radius equal to the contribution of both the protein and the heavy atoms and centre it at $-|F_H|$, it immediately becomes apparent that there are only two possible values for F_P:

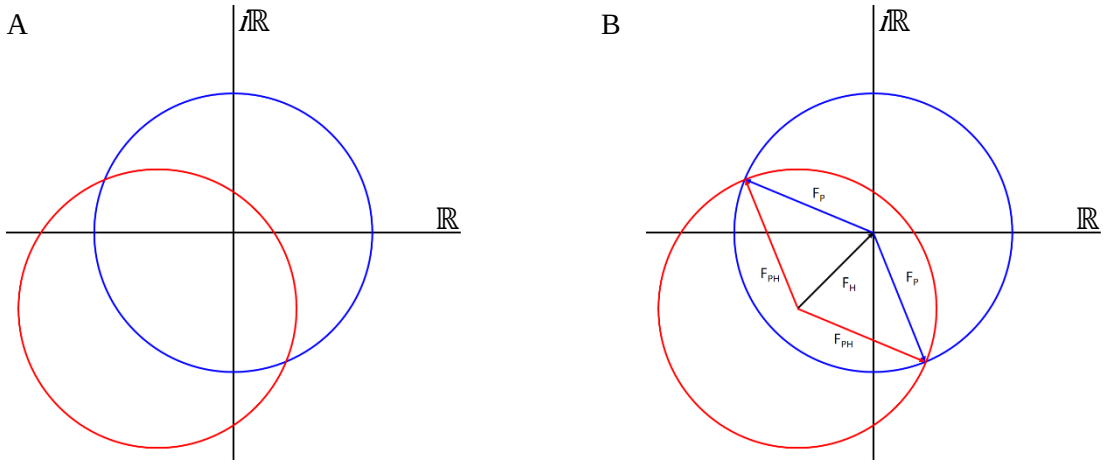


Figure 2.8 Phase contribution of native protein and heavy atoms. (A) Considering a heavy atom derivative gives us a second circle that crosses the first at two points. (B) From this, we are able to calculate the phase of the reflection in the native dataset.

If we take a third dataset with different heavy atoms, we are able to determine which of these values is correct (figure 2.9)

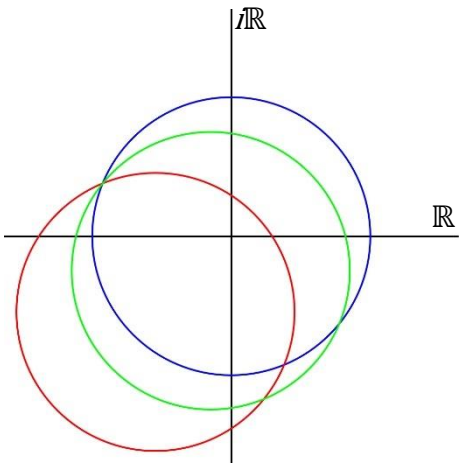


Figure 2.9 Adding a third derivative. With a third dataset, the single possible solution is revealed.

However, this assumes we have perfect data. In reality, all measurements will contain errors, resulting in errors in the phases, which, in our argand diagrams, "blurs" our circles (figure 2.10):

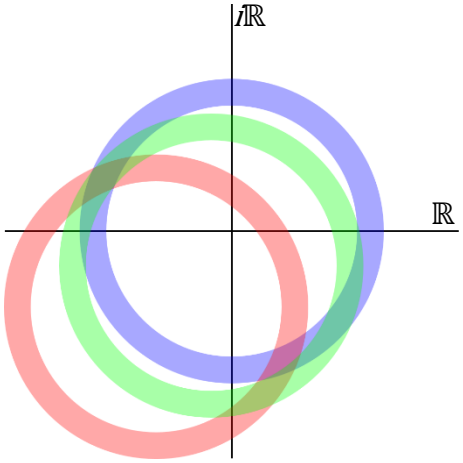


Figure 2.10 Errors in measurements. The presence of errors means that we are unable to exactly determine our amplitudes or our phases.

The probability distribution of the phases, and therefore, essentially, the likely degree of error, is described by Hendrickson-Lattman coefficients, A, B, C and D (Hendrickson & Lattman, 1970):

$$P(\varphi) = N^{A\cos\varphi + B\sin\varphi + C\cos 2\varphi + D\sin 2\varphi}$$

Friedel's law tells us that conjugate pairs of reflections (known as Friedel pairs) will have opposite phases and intensities of equal magnitude. This relationship is broken by the presence of an anomalous signal, resulting in slight changes to both the phase and intensity. Knowing this, by inspecting a Friedel pair, in a similar vein to how isomorphous replacement works, we are able to calculate the contribution to the reflections from anomalous diffraction.

The first structure to be solved by solved by single wavelength anomalous diffraction (SAD) was that of crambin (Hendrickson & Teeter, 1981), it was around another 20 years before the technique became more widely used due to advances in technology (Dauter *et al.*, 2002). In SAD, two potential phases can be determined due to the offset in the Friedel pairs caused by the anomalous signal:

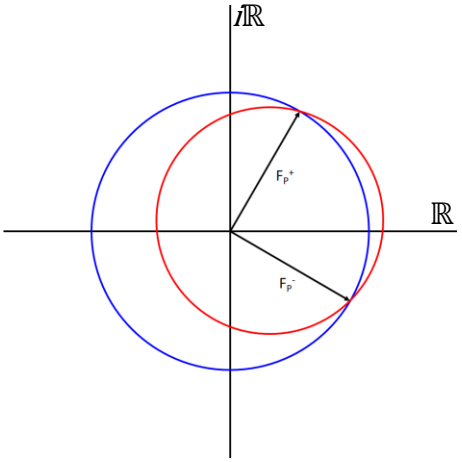


Figure 2.11 Phase contributions of Friedel pairs in SAD. Two possible phases can be determined, with the ambiguity solved through density modification.

The phase ambiguity resulting from the two possible solutions can be solved by density modification. Multi-wavelength anomalous diffraction (MAD) (Hendrickson, 1991) compares anomalous signals at a range of wavelengths around the absorption peak, allowing a single solution to be determined, as in figure 2.9.

The more accurate our measurements of the intensities are, the more precisely we are able to determine the phases, so the more defined our circles are. To achieve precise intensity measurements, it is necessary to collect high multiplicity data, for reasons given in section 2.3.3.2. SAD only uses a single dataset and exploits very slight differences in the probabilities of certain phases so requires much more precise measurements of the anomalous signal and therefore, much higher multiplicity (Dodson, 2003).

After initial phases have been determined, direct methods or Patterson methods can be used to place heavy atoms. The phases can then be gradually iteratively improved using density modification (Taylor, 2003). A range of methods of density modification are available, however the most commonly used method is probably solvent flattening. This uses the fact that very little solvent will be visible in the final structure as much of the solvent will be disordered. Because of this, any features that are seen in the map that are in regions predicted to be solvent are likely to be noise caused by the inaccurate phases. These features are weakened, while protein features are strengthened. Improved phases are then calculated from the new, more accurate, map. It is very likely that this map will now have improved protein density, from here, we can begin to build the protein structure (Cowtan & Zhang, 1999). This involves

iterative rounds of refinement and gradual tracing of the protein backbone as it appears. This can be done manually but is much faster when programs like *Buccaneer* (Cowtan, 2006) and ARP/wARP (Langer *et al.*, 2008) are used, which automate the process.

2.3.6 Completing a Structure

2.3.6.1 Refinement

Once a structure has been solved, it must be refined. The initial solution, although largely correct, is likely to have significant errors. A range of electron density maps can be calculated to identify errors in the model, although the F_O - F_C and $2F_O$ - F_C maps are the most common (Minichino *et al.*, 2003). F_O are the observed structure factors amplitudes, which are calculated from the measured intensities. These are combined with phases determined from the current model. While F_C are the structure factor amplitudes calculated from the current model, which are again combined with the calculated phases. A map calculated from F_O - F_C highlights differences between our model and our data and gives positive density where atoms are missing, negative density where there are atoms in the model that should not be there and no density when the model is correct. While a $2F_O$ - F_C map emphasises correct regions and also strengthens the weak density resulting from missing atoms. Programs like Coot (Emsley & Cowtan, 2004) can be used to manually adjust the model to account for the differences between F_O and F_C.

After a round of manual model building with Coot, the structure is fed into a refinement program such as REFMAC5, which uses maximum likelihood to adjust the model to fit the data and certain geometric restraints (Murshudov *et al.*, 2011). This process is repeated, iteratively improving the structure so that it reflects the experimental data, whilst preventing it from straying too far from established protein geometry (Rupp, 2010). With each round of refinement, the phases are recalculated based on the new model, resulting in better electron density, into which the structure can be further refined.

2.3.6.2 Validation

A range of tools are available during and after refinement of a structure to ensure that the structure accurately reflects the data and adheres to established geometry. Potentially the most well-known form of validation is the Ramachandran plot. In 1963, It was noted that only certain bond angles appear in peptide backbones and that some are more favoured than others. Occasionally, outliers are seen, but the vast majority of bond angles will conform to the angles permitted by the Ramachandran plot. By examining the bond angles present in a structure and comparing them to those known to occur, it can be determined whether the structure is likely to be valid (Ramachandran *et al.*, 1963). Although the precise allowed angles have varied slightly, and separate plots are now produced, at the very least, for proline and glycine, modern Ramachandran plots still closely resemble the original (Wlodawer, 2017). The Ramachandran plot considers two bond angles within the peptide backbone, however many other bonds exist in proteins and these, too should be considered when validating a structure (Morris *et al.*, 1992).

As well as bond angles, the bond lengths within a structure can also be validated. Over the years, large amounts of very high resolution data from small peptides has shown that all chemical bonds have ideal lengths and that the deviation from these lengths is minimal. Engh and Huber demonstrated how these bond lengths should be used in refinement of a protein structure (Engh & Huber, 1991). While there has been a level debate about exactly how closely a protein structure should conform to these ideals, all structures do conform and wild deviation should be considered a major "red flag" (Jaskolski *et al.*, 2007; Wlodawer *et al.*, 2008; Wlodawer, 2017).

Although the Ramachandran plot is arguably the most well-known form of protein structure validation, a different form of R-factor to those discussed earlier and a derivative of it, R_{free}, are arguably the most important. The R-factor, also known as the R_{cryst} or R_{work} calculates the sum of the differences between the observed structure factors and the calculated structure factors and presents it as a fraction of the sum of the observed structure factors (Rupp, 2010):

$$R = \frac{\sum\left||F_{obs}| - |F_{calc}|\right|}{\sum|F_{obs}|}$$

As the model becomes closer to the true structure of the protein, the calculated structure factors should approach the observed structure factors, resulting in the R-factor approaching zero. This does not happen, however, as the observed structure factors will always have a degree of error and a limited resolution, primarily due to a significant portion of the unit cell – the solvent – being disordered. It is possible though, to continue to reduce the R-factor of a structure without actually improving the structure, this can happen as a result of building into noise in the electron density map simply by adding more to the model, a practice known as "overfitting" (Rupp, 2010). To combat this, R_{free} was introduced (Brunger, 1992). R_{free} is calculated in the same way as the traditional R-factor, however, rather than using all reflections for refinement, a small portion of them are set aside (usually around 5% or 1000 reflections (Dodson *et al.*, 1996)). As the noise in data is random, building into this noise will not reduce R_{free}, so overfitting can be detected if $R - R_{free}$ is greater than approximately 0.07. Conversely, a difference of less than approximately 0.02 indicates that the free reflections may not truly be free and it will be likely that they have been used in a recent round of refinement, although the difference does tend to decrease as resolution improves (Wlodawer *et al.*, 2008).

Although exclusion of a large amount of experimental data only has a minor effect on the structure (Brunger, 1992), some resistance exists to using too many reflections in the free set. To combat this, the "free-kick" method was recently developed (Praznikar & Turk, 2014). This method adds a step of adjusting all coordinates in proportion to their approximate error. Structure factors are calculated from these "kicked" coordinates and are compared to the structure factors calculated from the model, which through a maximum likelihood function, allows the determination of R_{kick}, which is analogous to R_{free} in that it gives a representation of the degree of error

in the model and cannot be biased by over fitting, but it allows all data to be used in refinement. This method, however, has not yet gained widespread usage.

Refinement of a structure normally continues until there are no more features in the map still to be explained and multiple successive rounds of refinement with minor adjustments to the structure, such as addition or removal of marginal water molecules, have no effect on the R-factors (Rupp, 2010).

2.3.6.3 Data Depostion

Once refinement of a structure is complete, the structure is deposited in the protein data bank (PDB). The PDB is a freely accessible online repository containing the vast majority of published macromolecular structures over the last 30 years and a number of older structures. Submission has been considered mandatory by most journals since the early 1990s. This makes the structure available to the wider scientific community (Berman *et al.*, 2000).

2.4 Summary

All of the theory discussed here is much more complex and much more detailed than could possibly be covered within the scope of this thesis. However, this chapter serves as a brief overview of the theory behind the techniques used in chapters 4, 5, and 6.

Chapter 3

Methods

3.1 Plasmid manipulation

3.1.1 Transformation

Vectors were transformed into One Shot TOP10 (Thermo Fisher Scientific) or BL21-CodonPlus (Agilent Technologies) *Escherichia coli* competent cells, for plasmid propagation or expression respectively, using the heat shock method. 30 µl of pre-prepared competent cells had 1 µl DNA added at approximately 200 ng μl^{-1} while on ice. After 5 minutes, the cells were heat shocked at 42 °C for 90 seconds before returning to ice for 5 minutes. The cells then had 150 µl of lysogeny broth (LB – 1% (w/v) tryptone, 0.5% (w/v) yeast extract, 1% NaCl) (Bertani, 1951) added and were incubated with shaking at 37 °C for 1 hour. Cells were plated on LB Agar (LB with 1.5% agar) supplemented with appropriate antibiotics and incubated at 37 °C overnight.

A single colony was selected from the plates and used to inoculate 10 ml LB supplemented with appropriate antibiotics. This culture was grown overnight and either used to make glycerol stocks or for plasmid extraction.

3.1.2 Glycerol stock preparation

500 µl of overnight culture was added to 500 µl of 80% glycerol and mixed by inversion. Cells were flash frozen in liquid nitrogen and stored at -80 °C for later use to inoculate overnight cultures for expression or plasmid extraction.

3.1.3 Plasmid extraction

Expression plasmids were extracted from transformed TOP10 cells using a Wizard® Plus SV Miniprep kit (Promega) according to the manufacturer's protocol. A 10 ml overnight LB culture was centrifuged for 10 minutes at 5000 g. The supernatant was poured off and the cell pellet suspended in 250 µl cell resuspension solution. Cells were lysed by addition of 250 µl cell lysis solution, which was mixed by inversion and incubated at room temperature for 5 minutes. 10 µl alkaline protease solution was added to degrade released nucleases, which was also mixed by inversion and incubated at room temperature for 5 minutes. Alkaline proteases were deactivated by addition of 350 µl neutralisation solution which was mixed by inversion. The mixture was clarified by centrifugation at 13,200 g for 10 minutes.

Cleared lysate was transferred to a spin column and centrifuged at 13,200 g for 1 minute. The DNA bound to the spin column membrane was washed by addition of 750 µl column wash solution diluted with 95% ethanol and centrifuged for 1 minute, followed by addition of 250 µl diluted column wash solution and centrifugation for 2 minutes. DNA was dissolved in 50 µl nuclease free water and removed from the membrane by centrifugation into a 1.5 ml microcentrifuge tube. The concentration and purity of extracted DNA was assessed using a nanodrop 2000c (Thermo Fisher Scientific) before storing at -20 °C for sequencing or further use.

3.1.4 Sequencing

All constructs were sent for sequencing by Eurofins to confirm that the gene cloned into the expression vectors was correct. Sequencing results were translated using ExPASy Translate (Gasteiger *et al.*, 2003) and aligned to the known sequence with Clustal Omega (Sievers *et al.*, 2011).

3.2 Expression

3.2.1 Native Protein

10 ml overnight LB cultures supplemented with appropriate antibiotics were inoculated from transformed BL21-CodonPlus glycerol stocks and grown overnight at 37 °C with shaking. These were used to inoculate 500 ml LB or TB (terrific broth – 2.5% (w/v) yeast extract, 2% (w/v) tryptone, 0.5% (v/v) glycerol, 17 mM KH_2PO_4, 72 mM K_2HPO_4, pH 7.5) (Tartof & Hobbs, 1987) cultures supplemented with appropriate antibiotics. Cultures were grown to an OD_{600} of 0.6-0.8 before expression was induced by addition of 1 mM Isopropyl-β-D-1-thiogalactopyranoside (IPTG). Cultures were either cooled to 16 °C for overnight expression, or kept at 37 °C for 4-hour expression before harvesting.

Cell pellets were harvested by centrifugation at 8000 g for 10 minutes. The pellets were then washed by resuspension in appropriate lysis buffer before being centrifuged at 5000 g for 15 minutes and flash frozen in liquid nitrogen for storage at -80 °C.

3.2.2 Selenomethionine derivatives

For selenomethionine derivatives, 10 ml overnight LB cultures were grown as normal and centrifuged at 5000 g for 10 minutes. The supernatant was discarded and the pellet resuspended in 25 mM Tris, 200 mM NaCl, pH 8.0. This was centrifuged and resuspended again to remove as much LB as possible. The resuspended pellet was used to inoculate 500 ml selenomethionine medium (Molecular Dimensions) supplemented with appropriate antibiotics. Cultures were grown to an OD_{600} of approximately 1.0 before the temperature was reduced to 16 °C and methionine production was inhibited by addition of 100 μg ml^{-1} lysine, phenylalanine and threonine and 50 μg ml^{-1} leucine, isoleucine and valine (Walden, 2010). After 15 minutes, 60 μg ml^{-1} selenomethionine was added and expression was induced after a further 15 minutes by addition of 1 mM IPTG.

3.3 Purification

3.3.1 Solubly expressed protein

Pellets were resuspended in an appropriate lysis buffer and lysed at 20 KPSI (138 MPa) in a Constant Systems Cell Disruptor (French press). The lysate was centrifuged at 63,000g for 25 minutes. Proteins were initially purified from the cleared cell lysate by affinity chromatography using HiTrap columns (GE Life Sciences). If required, the protein was concentrated in a spin concentration with an appropriate molecular weight cut-off and was further purified by size exclusion chromatography using a Superdex 200 16/60 (GE Life Sciences) or buffer exchanged in to an appropriate buffer using a HiPrep 26/10 column (GE Life Sciences).

Efficacy of purification protocols was assessed by SDS-PAGE using Bis-tris gels. Protocols were optimised to improve purity and yield.

3.3.2 Inclusion body purification

Pellets were resuspended in lysis buffer (25 mM Tris, 200 mM NaCl, pH 8.0) and lysed using the same method as for soluble protein. Lysate was centrifuged at 16,000g for 25 minutes. Pellets were resuspended in lysis buffer with 1% Triton X-100.

Centrifugation was repeated and followed by a final wash step with lysis buffer and centrifugation.

Inclusion bodies were solubilised by resuspending the pellets in solubilisation buffer (8 M urea, 25 mM Tris, 150 mM NaCl, 50 mM $CaCl_2$, 50 mM glutathione (GSH), pH 8.0) overnight. The solubilised inclusion bodies were centrifuged at 63,000 g for 25 minutes and then refolded by serial dialysis into GSTrap loading buffer. (10 mM Na_2HPO_4, 2.7 mM KH_2PO_4, 140 mM NaCl, 10 mM KCl, 2 mM DTT, pH7.3). To reduce the chance of misfolding and prevent formation of near-insoluble calcium phosphate, a six step dialysis protocol was used with each step lasting at least four hours. The solubilised protein was dialysed in to the following buffers:

1) 4 M urea, 25 mM tris, 150 mM NaCl, 50 mM $CaCl_2$, pH 8.0

2) 25 mM tris, 150 mM NaCl, 50 mM $CaCl_2$, pH 8.0

3) 25 mM tris, 150 mM NaCl, pH 8.0

4) 25 mM tris, 150 mM NaCl, pH 8.0

5) 10 mM Na_2HPO_4, 2.7 mM KH_2PO_4, 140 mM NaCl, 10 mM KCl, pH7.3

6) 10 mM Na_2HPO_4, 2.7 mM KH_2PO_4, 140 mM NaCl, 10 mM KCl, 2 mM DTT, pH7.3

Refolded protein was centrifuged at 63,000g for 25 minutes before using the same chromatography protocol as for solubly expressed protein.

3.3.3 Polyhistidine-tagged proteins

Pellets containing His-tagged proteins were initially lysed into a suitable buffer containing 20 mM imidazole (eg. 25 mM Tris, 200 mM NaCl, 20 mM Imidazole, pH 8.0). Cleared cell lysate was loaded on to a HisTrap column (GE Life Sciences) before bound proteins were eluted with a gradient of increasing imidazole concentration to 500 mM.

The imidazole concentrations at which impurities and the target protein eluted were calculated and used to modify the concentrations used in a step-wise protocol to achieve optimal yield and purity. The concentration in the lysis buffer was increased

to reduce non-specific binding while not preventing His-tagged protein from binding. It was then increased to a concentration capable of eluting the tagged protein.

3.3.4 Glutathione S-transferase-tagged proteins

GST-tagged proteins were lysed or, after refolding, dialysed into GST loading buffer (10 mM Na_2HPO_4, 2.7 mM KH_2PO_4, 140 mM NaCl, 10 mM KCl, 2 mM DTT, pH7.3). Samples were then loaded on to a GSTrap column (GE healthcare). Bound protein was eluted with GST elution buffer (50 mM Tris, 10 mM GSH, pH 8.0).

The tag was cleaved by addition of an appropriate amount of GST tagged human *rhinovirus* 3C protease followed by overnight dialysis into cleavage buffer (50 mM Tris-HCl, 150 mM NaCl, 1 mM EDTA, 1 mM DTT, pH 7.0). The protease and GST tag were removed by a second round of chromatography, during which the target protein did not bind to the column.

3.3.5 Size exclusion chromatography (SEC)

Proteins purified by affinity chromatography were concentrated to a volume of approximately 1 ml for size exclusion using a GE Superdex 200 16/60 to remove minor impurities and buffer exchange the sample into a suitable buffer for later steps.

3.3.6 Desalting

If the sample was pure enough after affinity chromatography, it was concentrated to a volume of approximately 2.5 ml and buffer exchanged using a GE HiPrep 10/60 desalting column.

3.3.7 Concentrating

Proteins were concentrated using spin concentrators with a molecular weight cut off of no more than half of the mass of the protein of interest. Membranes were washed with the concentrated protein sample after concentrating to prevent loss of sample.

3.3.8 Polyacrylamide gel electrophoresis

Samples were taken during expression and purification for analysis by Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE). The samples were analysed using Bis-Tris gels, which have been shown be more reliable and result in better resolution than more conventional Tris-glycine gels (Graham *et al.*, 2005),

containing 8, 10, 12 or 15% polyacrylamide as appropriate. NEB Unstained broad range protein markers were used for all gels. Electrophoresis was conducted at 90 V for 15 minutes followed by 150 V for approximately 45 minutes, until the dye front had reached the bottom of the gel. Gels were stained with Coomassie brilliant blue.

3.3.9 Mass spectrometry

When first purified to an acceptable level, samples were transferred either by dialysis, SEC, or desalting to an aqueous solution of 0.1% acetic acid. If the sample did not degrade, it was characterised by electrospray ionisation mass spectrometry performed by the University of Bath Chemical Characterisation and Analysis Facility. This was used to confirm that the purified protein was the protein of interest or determine the level of selenomethionine incorporation.

3.4 Crystallographic studies

3.4.1 Crystallisation

An approximate optimum concentration for crystallisation was determined using a Hampton Research Pre-Crystallisation Test (PCT), a four condition hanging-drop experiment designed to give an indication of whether the current protein concentration is suitable for crystallisation. Initial PCTs used protein concentrated to around 10 mg ml^{-1}. Based upon the presence and appearance of precipitate in the PCT, proteins were either diluted or concentrated for further PCTs until a suitable concentration was determined.

A wide range of Crystallisation conditions were screened with 96-well Art Robbins Instruments (ARI) sitting-drop Intelli-Plates using an ARI Phoenix crystallisation robot and up to 14 Molecular Dimensions HT-96 screens. This allowed potentially thousands of conditions to be screened. The majority of crystals produced were used in diffraction experiments at Diamond Light Source. Any conditions that produced crystals that showed sub optimal diffraction were optimised either using the Phoenix or through larger scale 24 well hanging-drop plates.

3.4.2 X-ray data collection

The majority of diffraction data were collected on MX beamlines at Diamond Light Source. Crystals were initially tested for diffraction by collecting three images separated by rotations of 45°. These images allowed the assessment of diffraction quality of the crystal and an approximate maximum resolution to be determined. Test images were auto-indexed using Mosflm (Battye *et al.*, 2011) and EDNA (Incardona *et al.*, 2009). If diffraction quality appeared suitable, a full dataset was collected via strategies determined using the test images. Earlier datasets were collected using higher-dose strategies aimed at getting enough data with a relatively small oscillation of the crystals, while later datasets were collected with newer fast, low-dose strategies aimed at maximising multiplicity to reduce errors (Diederichs & Karplus, 2013; Pothineni *et al.*, 2014). Data were auto-processed with Fast_dp and Xia2 pipelines using a range of software (CCP4, 1994; Winter, 2010) to approximately determine the quality of each dataset collected.

3.4.3 Reprocessing of X-ray data

Datasets of sufficient quality were reprocessed using Xia2 and *DIALS* (Waterman *et al.*, 2016) and data were scaled and merged with *AIMLESS* (Evans & Murshudov, 2013). The graph of R_{merge} versus batch number was used to selectively remove ranges of poorer quality images to improve the quality of the dataset. Dataset quality was determined based on a range of statistics detailed in Chapter 2, primarily based upon R_{merge} for earlier datasets and $CC_{1/2}$ for later datasets.

3.4.4 Structure determination

Once optimal statistics had been achieved using *AIMLESS*, structure solution via molecular replacement was attempted using Phaser (McCoy *et al.*, 2007) with models generated using a range of methods. This regularly proved unsuccessful, so experimental phasing or more complex molecular replacement methods were employed, detailed in the methods sections of respective chapters.

3.4.5 Refinement and validation

Structures were refined using successive rounds of REFMAC5 (Murshudov *et al.*, 2011) with manual model building with Coot (Emsley & Cowtan, 2004). This process

continued until the R_{work} and R_{free} ceased to decrease between successive rounds. A range of validation methods available in Coot and REFMAC5 detailed in chapter 2 were used throughout refinement and structures were checked towards the end of refinement with MolProbity (Chen *et al.*, 2010). Validation tools provided by the Protein Data Bank (PDB) were used before deposition of structure with the PDB (Berman *et al.*, 2000).

Chapter 4

Cwp84

4.1 Introduction

The *cwp84* gene is located in the s*lpA* locus six genes downstream from *slpA* itself (Karjalainen *et al.*, 2001; Sebaihia *et al.*, 2006; Monot *et al.*, 2011). The gene was first identified in 2001 and named after the approximately 84 kDa 803 residue protein it codes for, which is predicted to contain an N-terminal cysteine protease domain (Karjalainen *et al.*, 2001). Predictions show that this domain is followed by a region of approximately 170 residues of unknown structure and function and the three C-terminal cell wall binding domains (Eddy, 2008; Jones *et al.*, 2014; Finn *et al.*, 2016). Proteolytic activity was initially demonstrated against gelatine and the extra cellular matrix proteins fibronectin, laminin, and vitronectin, however it was shown to be unable to cleave type IV collagen (Janoir *et al.*, 2004; Janoir *et al.*, 2007).

Although the majority of Cwps are upregulated in the presence of sub-MIC (minimum inhibitory concentration) ampicillin, clindamycin and metronidazole, presumably as a stress response, Cwp84 is one of only four Cwps that is upregulated in the presence of sub-MIC amoxicillin (Emerson *et al.*, 2008).

Cwp84 has been determined to be responsible for the cleavage of SlpA to form HMW SLP and LMW SLP (Karjalainen *et al.*, 2001; Kirby *et al.*, 2009; Dang *et al.*, 2010). Cwp84 knockouts present full length SlpA on the surface of the cell. This results in an abnormal S-layer and the presence of SlpA, Cwp2 and Cwp66 in growth media, which is not seen in the wild type (Kirby *et al.*, 2009). Knockouts also show aberrant colony morphology, grow at half their usual rate, and have a propensity to aggregate (Kirby *et al.*, 2009; de la Riva *et al.*, 2011). A Cwp84 knockout strain was, however, still able to cause *C. difficile* infection (CDI) in hamsters (Kirby *et al.*, 2009) and inhibition of Cwp84 did not prevent growth *in vitro* (Dang *et al.*, 2010). Despite this, It has been suggested that perturbation of S-layer formation may make the bacterium more susceptible to antibiotics (Dang *et al.*, 2010).

This information shows that Cwp84 is an important component of the S-layer that should be studied further. Cysteine proteases are ubiquitous enzymes and are involved in a wide range of processes (Toh *et al.*, 2010). Because of this, a good understanding of the subtle structural and mechanistic differences between Cwp84

and other cysteine proteases is required to fully elucidate its function within the S-layer if it is to be used as a future drug target.

4.1.1 Papain proteases

Cwp84 is specifically a C1A cysteine protease. The first of these to be identified, and therefore the defining member of the family, was papain, which was isolated from *Carica papaya*. Its activity was noted to be similar to that of trypsin (Wurtz & Bouchut, 1879; Martin, 1885). The structure of papain was determined in 1968 and was shown to be formed by two lobes, an N-terminal lobe comprised of α-helices and loops and a C-terminal lobe primarily formed by a β-sheet (figure 4.1) The catalytic dyad, a cysteine and a histidine, were shown to sit in the active site groove between the two lobes (Drenth *et al.*, 1968). Members of the family are synthesised with a propeptide that sits in the active site groove in the opposite orientation to that of the substrate,

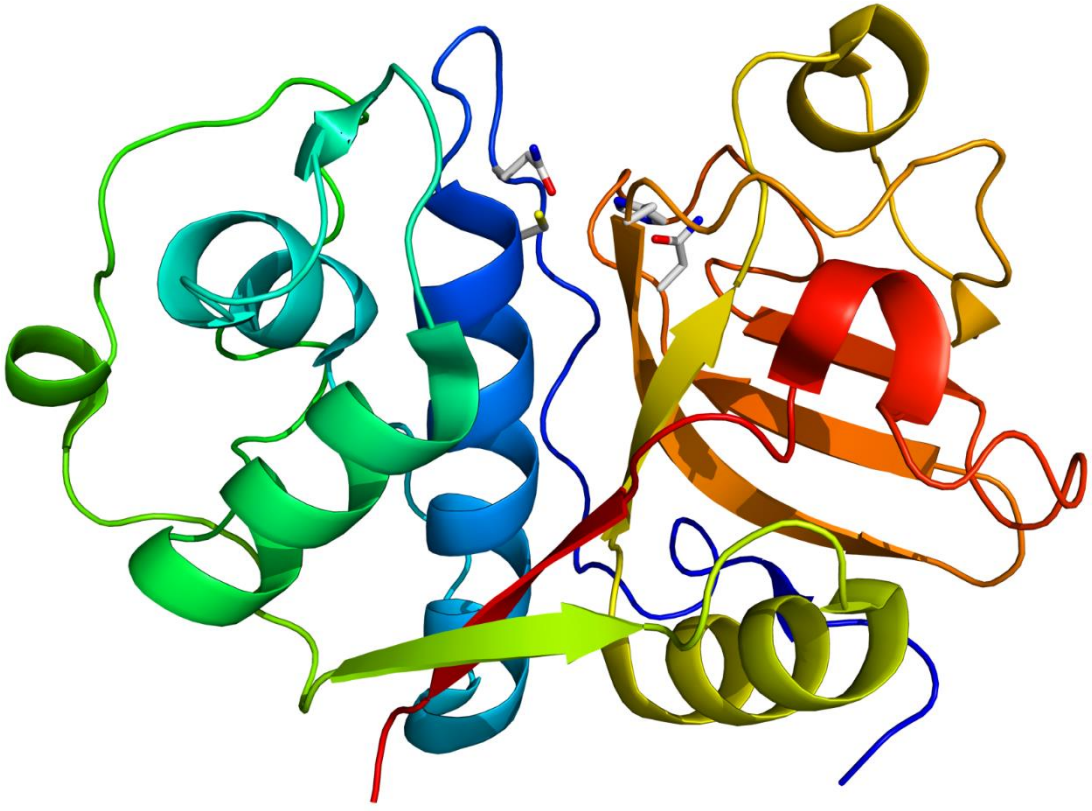


Figure 4.1 The structure of mature papain. The N-terminal lobe can be seen on the left and the C-terminal lobe on the right. The active site cleft is visible in the centre with the catalytic cysteine and histidine residues and assisting glutamate and aspartate. Produced from 1PPN (Pickersgill *et al.*, 1992). The ribbon is coloured from the N-terminus to the C-terminus, blue to red. The figure was generated with PyMol.

preventing cleavage (Schaschke *et al.*, 1998). The propeptide is important for correct folding of the enzyme and inhibits activity of the protease until it is cleaved (Fox *et al.*, 1992; Wiederanders, 2003).

Several cathepsins are members of the C1A cysteine protease family and are arguably the most studied of the group due to their association with a range of physiological disorders and potential use as drug targets (Hook *et al.*, 2015; Siklos *et al.*, 2015; Stoka *et al.*, 2016; Wen *et al.*, 2016). C1A cysteine proteases are generally classified as either cathepsin L-like or cathepsin B-like (figure 4.2) (Toh *et al.*, 2010). The two can be differentiated through the presence of a so-called "occluding loop" in cathepsin B-like cysteine proteases and differences in propeptide length (Sajid & McKerrow, 2002).

The occluding loop of cathepsin B-like cysteine proteases partially covers the S' end of the active site, that is, the region to which the residues of the substrate

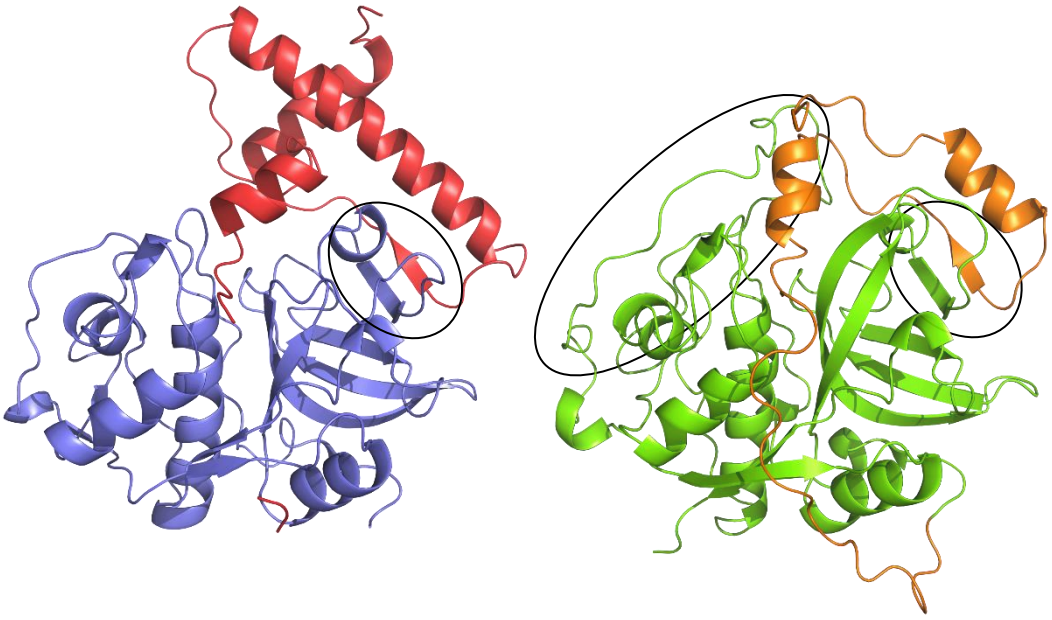


Figure 4.2 The structures of procathepsin K (left) and procathepsin B (right). Propeptides are coloured separately to the rest of the protein. The two lobe structure seen in papain is present. The propeptide sits in the active site groove (poor density for cathepsin K) and wraps around the protein. Cathepsin K, a cathepsin L-like cysteine protease possesses a significantly larger propeptide and lacks an occluding loop, which is circled on cathepsin B, as are the prosegment binding loops of the two proteins. Generated from 1PBH (Turk *et al.*, 1996) and 7PCK (Sivaraman *et al.*, 1999).

immediately following the scissile bond bind. This has two effects: firstly, cathepsin B-like cysteine proteases generally have greater substrate specificity than cathepsin L-like cysteine proteases as the occluding loop is able to block the entry of many would-be substrates into the active site. Secondly, while the protein retains endopeptidase activity, the occluding loop also confers carboxypeptidase activity to the enzyme – the ability to cleave residues from the C-terminus of the protein. A conserved HH motif within the occluding loop (His110, His111 in human cathepsin B) facilitates the binding of the C-terminus of substrates to the active site allowing the two C-terminal residues to be cleaved (Redzynia *et al.*, 2008; Tsuji *et al.*, 2008; Renko *et al.*, 2010).

As previously mentioned, the propeptide of C1A cysteine proteases sits in the active site of the protein in the opposite orientation to that of the substrate (Wiederanders, 2003). This region is the most proximal to the N-terminus of the mature enzyme. Moving towards the N-terminus of the propeptide, it wraps around the mature protein and forms a small two-stranded β-sheet with a part of the β lobe known as the prosegment binding loop (PBL). This is preceded in cathepsin B-like proteins by a short N-terminal helix and in cathepsin L-like proteins by a considerably longer helix which is itself preceded by a shorter N-terminal helix that folds back on the former (Coulombe *et al.*, 1996; Sivaraman *et al.*, 1999). Papain proteases are frequently, but not always (Nagler *et al.*, 1999; Dahl *et al.*, 2001), able to autoactivate (ChapetónMontes *et al.*, 2011; Beton *et al.*, 2012). Cwp84 has been shown to possess a 7 kDa propeptide and is unlikely to be capable of autoactivation, although the reason for this is currently unclear (de la Riva *et al.*, 2011).

4.1.1.1 Catalytic mechanism

The nomenclature of substrate residues and binding pockets within cysteine proteases follows a system that was first proposed 50 years ago (Schechter & Berger, 1967) . The residue immediately before the scissile bond is referred to as the P_1 residue with residues before that having increasing numbers based on their distance from the scissile bond. Residues after the scissile bond are referred to as P'_n residues. Their respective binding pockets are referred to as S_n and S'_n (Schechter & Berger,

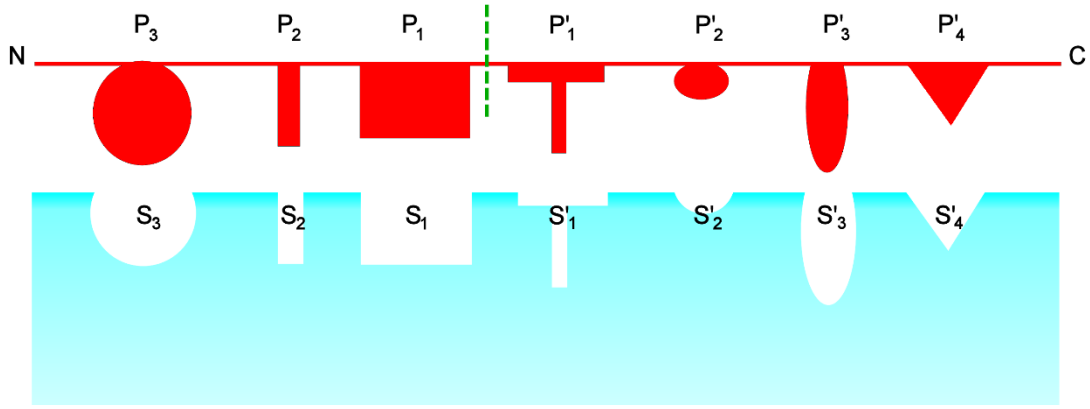


Figure 4.3 Protease nomenclature schematic. The scissile peptide is shown in red and the protease in cyan. The scissile bond is indicated by a vertical green line. Residues before the scissile bond are referred to as P_n, with n being determined based on their distance from the scissile bond, while residues after the scissile bond are referred to as P'_n. The respective binding pockets are referred to as S_n and S'_n.

1967; Sajid & McKerrow, 2002) (figure 4.3). The majority of interactions between cysteine proteases and their substrates and inhibitors occur between the P end of the substrate and S binding pockets (Matsumoto *et al.*, 1999). A significant portion of substrate specificity is determined by the S_2 pocket, to which the P_2 residue of the substrate binds. Specifically, a residue forming the outermost surface of the pocket frequently interacts with the P_2 residue of the substrate. In papain this residue is a serine, in cathepsin L it is an alanine and in cathepsin B it is a glutamate (Sajid & McKerrow, 2002).

The two main catalytic residues of C1A cysteine proteases, a cysteine and a histidine, are frequently assisted by a glutamine and/or an asparagine. These residues have been identified in Cwp84 as Gln110, Cys116, His262, and Asn294 (Savariau-Lacomme *et al.*, 2003; Dang *et al.*, 2010; de la Riva *et al.*, 2011). The histidine, which may be polarised by the asparagine, deprotonates the cysteine which nucleophilically attacks the carbonyl carbon of the scissile bond. The transition state is stabilised by an oxyanion hole formed by the glutamine. The histidine then protonates the scissile nitrogen, allowing the P' end of the substrate to leave the active site, while the P end forms an acyl-enzyme complex. The acyl-enzyme complex is hydrolysed, with the OH

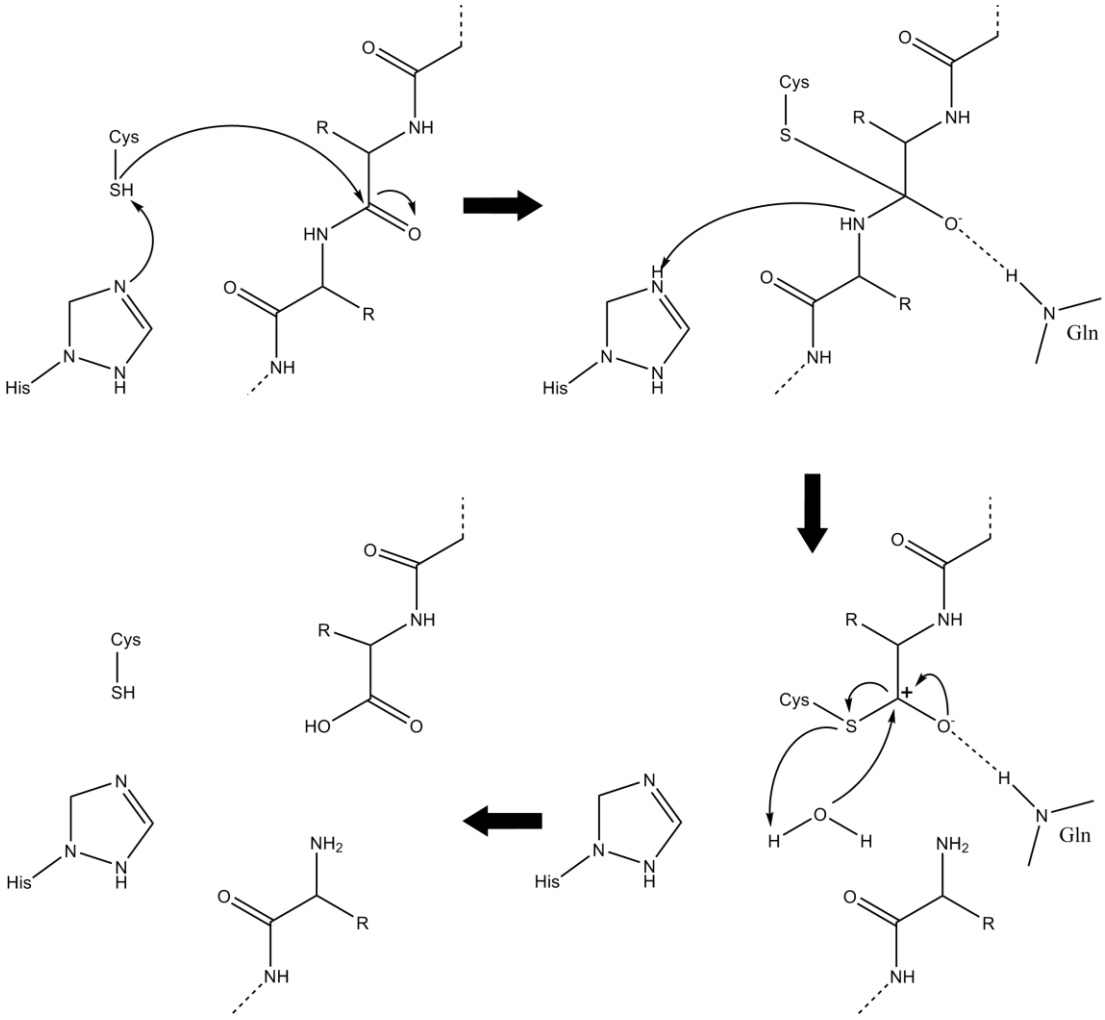


Figure 4.4 The mechanism of C1A cysteine proteases. Starting in the top left, the thiolate group of a deprotonated cysteine nucleophilically attacks the carbonyl carbon of the scissile bond while the P' leaving group is protonated by the histidine and the transition state is stabilised by the glutamine. The transition state is ultimately hydrolysed and the cysteine reprotonated.

group being added to the P end of the substrate and the remaining proton is added to the cysteine (figure 4.4) (Toh *et al.*, 2010).

4.1.1.2 Inhibition

As cysteine proteases have long been considered important factors in a wide range of physiological disorders and microbiological infections, they are also considered to be important drug targets, however, they also have a large number of important roles in cellular functions (Sajid & McKerrow, 2002; Hook *et al.*, 2015; Siklos *et al.*, 2015; Stoka *et al.*, 2016; Wen *et al.*, 2016). Naturally then, if cysteine proteases are to be

used as drug targets, a good understanding of the subtle differences between different cysteine proteases, their mechanisms and their interactions with inhibitors is required. This will allow the design of new, more specific inhibitors, which will retain their effect at lower doses and reduce off-target side effects.

Two starting points for the design of cysteine protease inhibitors are common. As the propeptide sits in the active site in the reverse direction to that of the substrate, it already fits well and cannot be cleaved. This, in effect, makes it a natural inhibitor. Many natural inhibitors of cysteine proteases therefore mimic the binding of the propeptide. Alternatively, E-64 (L-transepoxysuccinyl-L-leucyl-1-amino-4-guanidino-butane), which was first isolated from *Aspergillus japonicas* by Hanada *et al.* (1978), is a relatively non-specific "go-to" cysteine protease inhibitor (Barrett *et al.*, 1982). Many attempts at designing cysteine protease inhibitors therefore use either the propeptide or E-64 as a starting point.

Before activation, cysteine protease propeptides sit in the active site, preventing catalysis. After cleavage, propeptides are still able to act as inhibitors with K_is in the nanomolar range. Propeptides are sometimes able to inhibit cysteine proteases other than the one that they originated from but with affinities orders of magnitude lower (Yamamoto *et al.*, 2002; Wiederanders *et al.*, 2003). The frequent mimicry of the P_2 residue of the substrate by propeptides and the substrate selectivity conferred by the S_2 subsite can also be exploited by inhibitors (Coulombe *et al.*, 1996; Sivaraman *et al.*, 1999).

As the systematic name indicates, E-64 possesses a transepoxysuccinyl group amide bonded to an L-leucyl group, which is amide bonded to the amine end of a 1-amino-4-guanidino-butane group (figure 4.5). E-64 binds to cysteine proteases irreversibly when the catalytic cysteine nucleophilically attacks carbon-2, part of the epoxy ring. The resulting carboxylate is stabilised by the oxyanion hole and the catalytic histidine. The remaining interactions with the leucine residue and aminoguanidinobutane group can vary significantly, based on the structure of the specific active site, so these regions of E-64 are frequently subject to variation to produce better inhibitors (Matsumoto *et al.*, 1999).

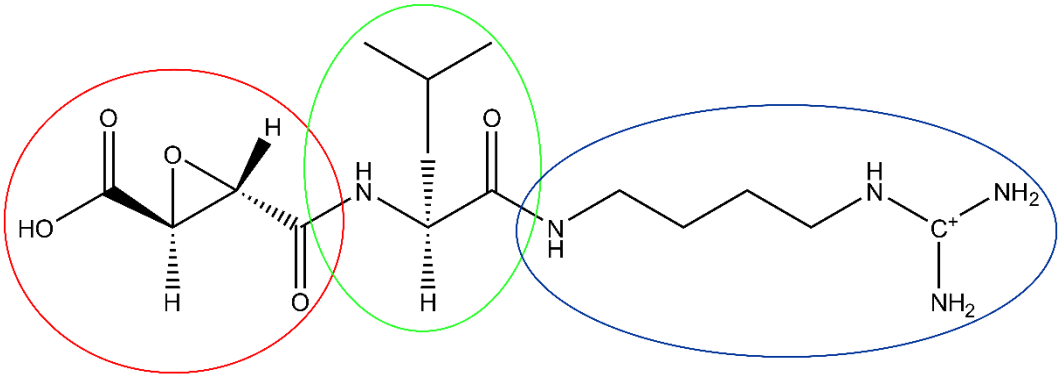


Figure 4.5 The structure of E-64 – The expoxysuccinyl group is circled in red, L-leucyl in green and aminoguanidinobutane in blue. The majority of conserved interactions with cysteine proteases are through the epoxysuccinyl group, while interaction with the rest of the structure can vary so these regions are modified to make more potent inhibitors.

4.2 Methods

4.2.1 Preliminary work

A synthetic construct coding for full length Cwp84 was ordered from Geneart by Jon Kirby, who, along with Chris Chambers, cloned the gene and various truncated or mutated versions of it into a range of different expression vectors. Many of the constructs showed poor expression or degradation. Ultimately, useable expression was seen for a construct cloned into the GST tagged vector pGEX-6P-1 coding for residues 33-497, lacking the three cell wall binding domains with an active site C116A mutation.

As well as this, six other constructs were produced, all with the same C116A mutation and possessing either the portions coded for by the earlier construct and one, two, or three cell wall binding domains or just one, two, or three cell wall binding domains on their own (figure 4.6).

4.2.2 Expression and purification

Construct 1 was expressed at 16 °C overnight according to the native expression protocol given in Chapter 3 and purified using the GST purification protocol. After the second round of GSTrap purification to remove the GST tag, the protein was further purified by size exclusion chromatography into 50 mM Tris pH 8.0. A

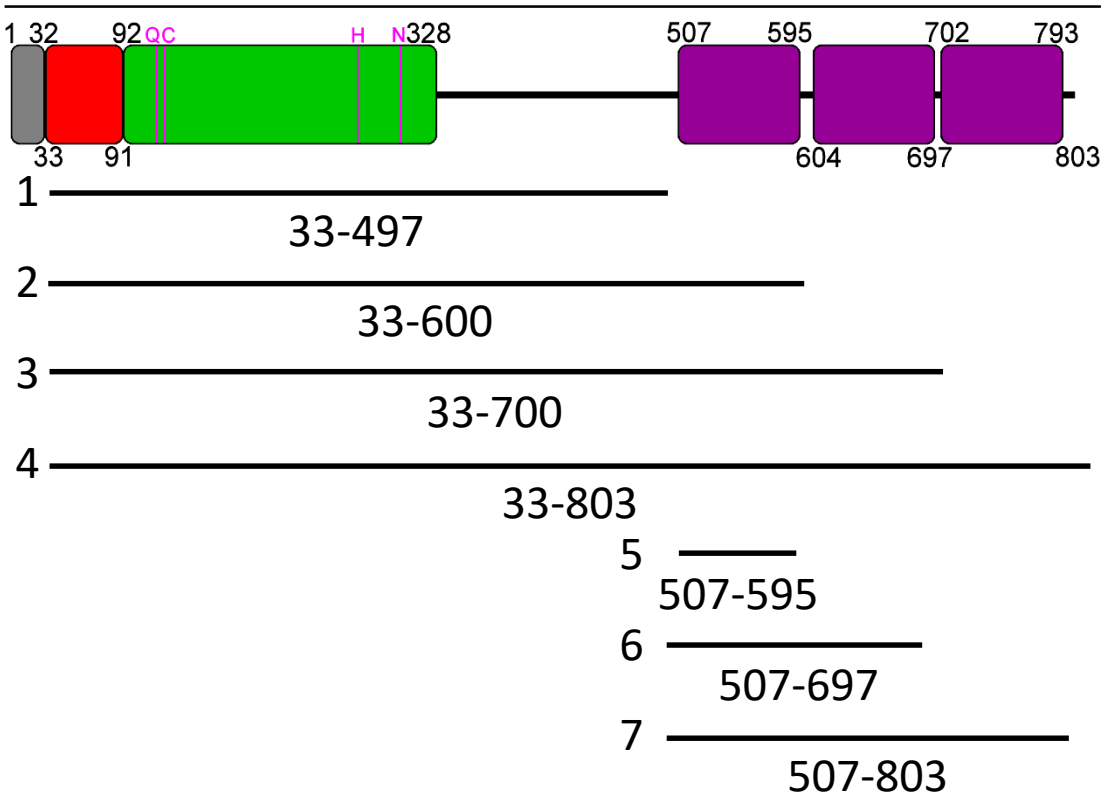


Figure 4.6 Cwp84 constructs – The predicted domain structure of Cwp84 is shown, the signal peptide is shown in grey, the propeptide in red, the cysteine protease domain in green with predicted active site residues in pink and the cell wall binding domains are shown in purple. Seven constructs were ordered from Geneart, numbered 1-7 and cloned into pGEX-6P-1 by Jon Kirby and Chris Chambers. All constructs possessed an active site C116A mutation.

selenomethionine derivative was expressed according to the selenomethionine expression protocol given in Chapter 3 and purified and crystallised in the same way as the native protein with the inclusion of 5 mM DDT in all buffers.

Constructs 2 to 7 were expressed for 4 hours at 37 °C. Constructs 2 to 4 were purified according to the inclusion body and GST purification protocols followed by size exclusion. Constructs 5 to 7 were purified according to the GST purification protocol followed by size exclusion.

4.2.3 Propeptide cleavage

It has previously been noted that the propeptide of Cwp84 can be removed by incubation with trypsin (Janoir *et al.*, 2007; de la Riva *et al.*, 2011). A protocol was developed based on this. Eluate from the first round of GST purification of construct

1, before cleavage of the GST tag, was concentrated to a volume of approximately 1 ml. The protein was incubated with trypsin in 20mM $CaCl_2$ at an approximate molar ratio of 10:1, $Cwp84_{33\text{-}497\_C116A}$:trypsin, at 37 °C for 45 minutes, this resulted in removal and degradation of the GST tag and propeptide, generating $Cwp84_{92\text{-}497\_C116A}$. After cleavage, the protein was centrifuged at 13,200 g at 4 °C for 10 minutes before mature protein was separated from degraded peptides by size exclusion chromatography on a Superdex 16/60 (GE Healthcare).

The mass of all three proteins was determined by mass spectrometry after purification as described in Chapter 3. An N-terminal sequence of SSVAY, corresponding to residues 92 onwards was confirmed by N-terminal sequencing (Alta Bioscience).

4.2.4 Crystallisation of construct 1

Optimal concentrations for crystallisation of 10-20 mg ml^{-1} for construct 1 with the propeptide (residues 33-497) and 15-25 mg ml^{-1} for construct 1 after propeptide cleavage (residues 92-497) were determined by PCTs and crystallisation conditions were determined via screening as described in Chapter 3.

4.2.6 Co-crystallisation of construct 1

Construct 1 without the propeptide was incubated on ice overnight with two peptides based on the cleavage site in SlpA at a molar concentration of 10:1, peptide:protein. The protein was also incubated overnight with the cysteine protease inhibitor E-64 at a molar concentration of 5:1 inhibitor:protein. Crystallisation screens were set up in the usual way.

4.2.6 X-ray data collection and processing of construct 1 data

Crystals of construct 1 with the propeptide were observed in Molecular Dimensions Structure Screen 1&2 condition D7 (0.2 M ammonium sulphate, 30% PEG 4,000). These crystals were reproduced with a range of drop volumes. The crystals were cryoprotected by addition of 50% glycerol to a final concentration of approximately 30%. Crystals were flash cooled by placing in a 100 K nitrogen cryostream. A high resolution dataset was collected from one crystal on beamline I02 at Diamond Light

Source. The data were autoprocessed with Xia2 (Winter, 2010) pipeline 3d using *XDS* (Kabsch, 2010). The data were scaled with *AIMLESS* (Evans & Murshudov, 2013) but attempts at molecular replacement were unsuccessful.

Selenomethionine crystals grown in the same conditions were similarly cryoprotected with 30% glycerol. Crystals were flash cooled with liquid nitrogen and MAD data were collected from four crystals on beamline I04 at Diamond Light Source. Fluorescence scans were performed to determine the precise Se-K edge, calculated by *CHOOCH* (Evans & Pettifer, 2001), around which data were collected. Peak data were collected at 12,660 eV, inflection at 12,656 eV, high-energy remote data at 12,770 eV and low-energy remote at 12,550 eV. Peak and inflection data were collected from all four crystals with either a high- or low-remote dataset also collected from each for a total of two high- and two low-remote datasets.

As with the high resolution data, MAD data were autoprocessed with Xia2 (Winter, 2010) pipeline 3d using *XDS* (Kabsch, 2010). The data were scaled with *SCALA* (Evans, 2006), combined with CAD (Winn *et al.*, 2011) and put into the CRANK MAD pipeline (Ness *et al.*, 2004) using SCALEIT (Howell & Smith, 1992), AFRO (CCP4, 1994), CRUNCH2 (de Graaff *et al.*, 2001), BP3 (Pannu *et al.*, 2003; Pannu & Read, 2004), SOLOMON (Abrahams & Leslie, 1996) and 500 cycles of *Buccaneer* (Cowtan, 2006) and REFMAC5 (Murshudov *et al.*, 2011). Further model building and refinement were performed with *Coot* (Emsley & Cowtan, 2004) and REFMAC5.

Crystals of construct 1 without the propeptide were observed in PACT condition D7 (0.2 M NaCl, 0.1 M Tris pH 8.0, 20% PEG 6000) and MIDAS condition F7 (20% dimethyl sulfoxide, 20% Jeffamine M-2070) (Molecular Dimensions). The former condition was optimized by addition of 10% Silver Bullets condition E9 (0.2% 1,4-diamino-butane, 0.2% cystamine dihydrochloride, 0.2% diloxanide furoate, 0.2% sarcosine, 0.2% spermine, 20 mM sodium HEPES pH 6.8; Hampton Research). Both crystals were cryoprotected by addition of 1 µl 50% reservoir solution and 25% glycerol. The crystals were flash cooled with liquid nitrogen and data were collected on beamline I04 at Diamond Light Source. Data were autoprocessed with Xia2 pipeline 3dii and *XDS* and scaled with *AIMLESS*. Molecular replacement was performed with Phaser

(McCoy *et al.*, 2007) using the structure with the propeptide as a model but with the propeptide removed. Rounds of model building and refinement were performed with *Coot* and REFMAC5. All structures were validated using methods detailed in Chapters 2 and 3 before deposition in the PDB.

4.2.7 Lectin-like domain ELISA

Purified protein from construct 1 without the propeptide ($Cwp84_{92\text{-}497\_C116A}$) was sent to ThermoFisher for the production of rabbit polyclonal antibodies. IgGs were purified from a 56-day bleed using a protein A column. Purified Cwp84 from construct 1, both with and without the propeptide and concanavalin A (Sigma Aldrich), were diluted to 5 μg ml^{-1}, bound to a 96 well plate, and incubated for one hour with α-$Cwp84_{33\text{-}497\_C116A}$ or $Cwp84_{92\text{-}497\_C116A}$ and α-concanavalin A (Sigma Aldrich) at a range of concentrations to determine appropriate dilutions for use of the antibodies. The plates were washed three times with FBS incubated for one hour with HRP conjugated goat-antirabbit IgGs (Sigma Aldrich) washed with FBS again and developed by addition of TMB followed by stop solution (containing sulphuric acid). The oxidation of TMB was measured on a plate reader at 450 nm.

Glucose, mannose, galactose, N-acetylglucosamine, arabinose and maltose were bound to Nunc Maxisorp 96 well plates according to the manufacturer's instructions at 500 mM and 100 mM. The plates were washed with carbohydrate free blocking solution. Cwp84 and concanavalin A were incubated at 6 μM for one hour before a second wash step. The plates were incubated with α-Cwp84 at 0.6 μg ml^{-1} and α-concanavalin A at 2 μg ml^{-1} and washed again, incubated with HRP-antirabbit, washed and developed.

4.3 Results

4.3.1 Expression and purification with the propeptide

When expressed in LB medium, construct 1 was observed to produce some inclusion bodies however, enough protein was observed in the soluble fraction to proceed, discarding the insoluble fraction.

The protein could be purified to a high degree using a three step process as described in Chapter 3. The majority of purification was performed in the initial GSTrap step, which was followed by 3C protease cleavage of the GST tag and a second GSTrap step. The $Cwp84_{33\text{-}497\_C116A}$ bound to the column in the first step, allowing the removal of the majority of contaminants and did not bind in the second step, after tag removal. Therefore, except for a very small number of contaminants, $Cwp84_{33\text{-}497\_C116A}$ was the only protein present in the flow through from the second GSTrap step (figure 4.7A). Size exclusion chromatography was used to remove residual contaminants (figure 4.7B). The identity of the purified protein was determined by Electro-spray ionisation mass spectrometry (figure 4.8).

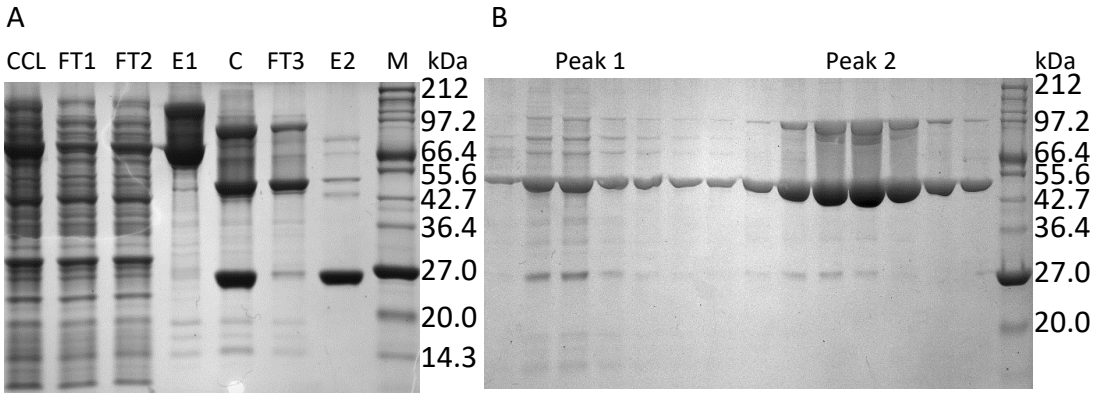


Figure 4.7 SDS-PAGE showing purification of $Cwp84_{33\text{-}497\_C116A}$. $Cwp84_{33\text{-}497\_C116A}$ with GST tag – 78 kDa, $Cwp84_{33\text{-}497\_C116A}$ without GST tag – 52 kDa, GST – 26 kDa. **(A)** GSTrap purification. CCL – cleared cell lysate. FT1 – early flow through from first GSTrap step. FT2 – late flow through from first GSTrap step. E1 – Eluate from first GSTrap step. C – 3C protease cleaved $Cwp84_{33\text{-}497\_C116A}$ and GST. FT3 – flow through from second GSTrap step, including $Cwp84_{33\text{-}497\_C116A}$. E2 – eluate from second GSTrap step, including GST tag. GST tagged $Cwp84_{33\text{-}497\_C116A}$ can be seen in the first four lanes. This is cleaved to produce $Cwp84_{33\text{-}497\_C116A}$ and GST, which are separated by a second GSTrap step. **(B)** Size exclusion chromatography. Two peaks were observed, the first was at the void volume of the column and contained $Cwp84_{33\text{-}497\_C116A}$ and significant amount of other contaminants while the second contained more $Cwp84_{33\text{-}497\_C116A}$ and a significantly smaller number of contaminants. Gels from $Cwp84_{33\text{-}497\_C116A}$ purifications frequently contained a band with a mass approximately double that of $Cwp84_{33\text{-}497\_C116A}$.

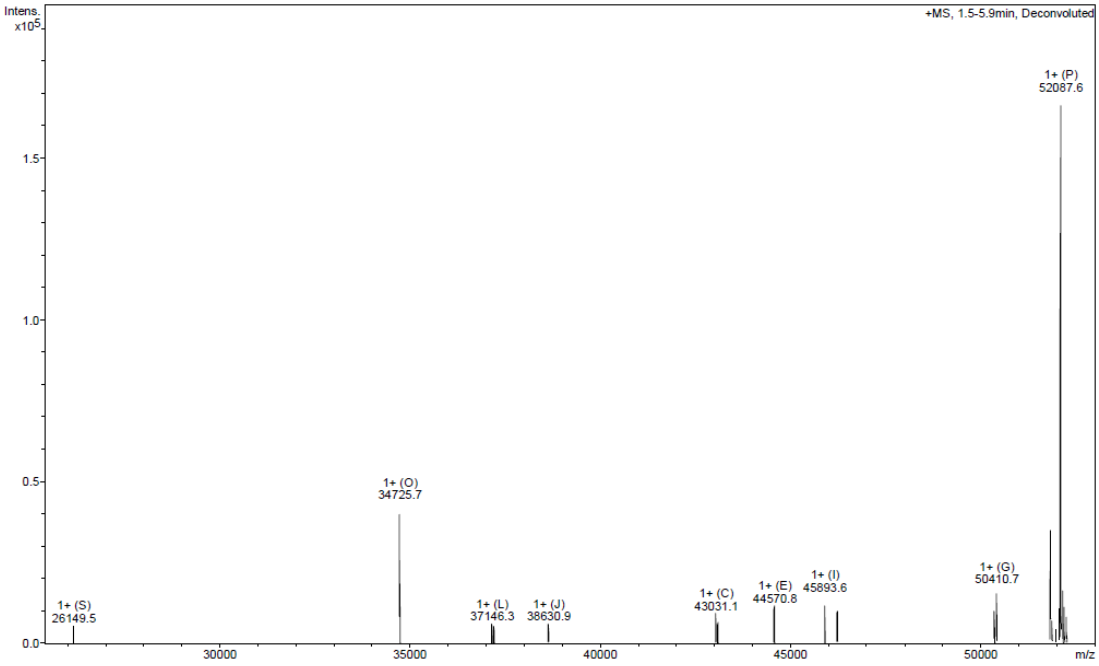


Figure 4.8 Mass spectrum of Cwp84$_{33\text{-}497_C116A}$. The peak at 52 kDa, which corresponds with the predicted mass of Cwp84 construct 1 is by far the most abundant species. This confirms that the expressed and purified protein is Cwp84$_{33\text{-}497_C116A}$, allowing it to be taken forward for crystallisation.

4.3.2 Purification without the propeptide

Cwp84$_{33\text{-}497_C116A}$ from construct 1 was initially expressed and purified with the propeptide in the same way. The protein was incubated with trypsin after the first GSTrap step (lane E1 in figure 4.7A) as described in Chapter 3 and products were separated by size exclusion chromatography. This was initially performed with 25 mM MOPS pH 7.0, which produced two peaks at around 80 ml. In an attempt to separate the peaks, the buffer used for this step was modified. Ionic strength of the buffer was found to have the greatest effect on peak separation: increasing the ionic strength of the buffer resulted in the peaks moving closer together, while decreasing it resulted in improved separation (figure 4.9). This is likely to have been a result of hydrophobic interactions with the Superdex medium (Tom Hutchison, GE Healthcare, Personal communication). Products from the tryptic digest were ultimately separated with 10 mM Tris, pH 8.0. The first peak was likely to have contained GST, while the second peak contained Cwp84$_{92\text{-}497_C116A}$ (figure 4.10). The identity of this product

was confirmed with electrospray ionisation mass spectrometry (figure 4.11) and N-terminal sequencing.

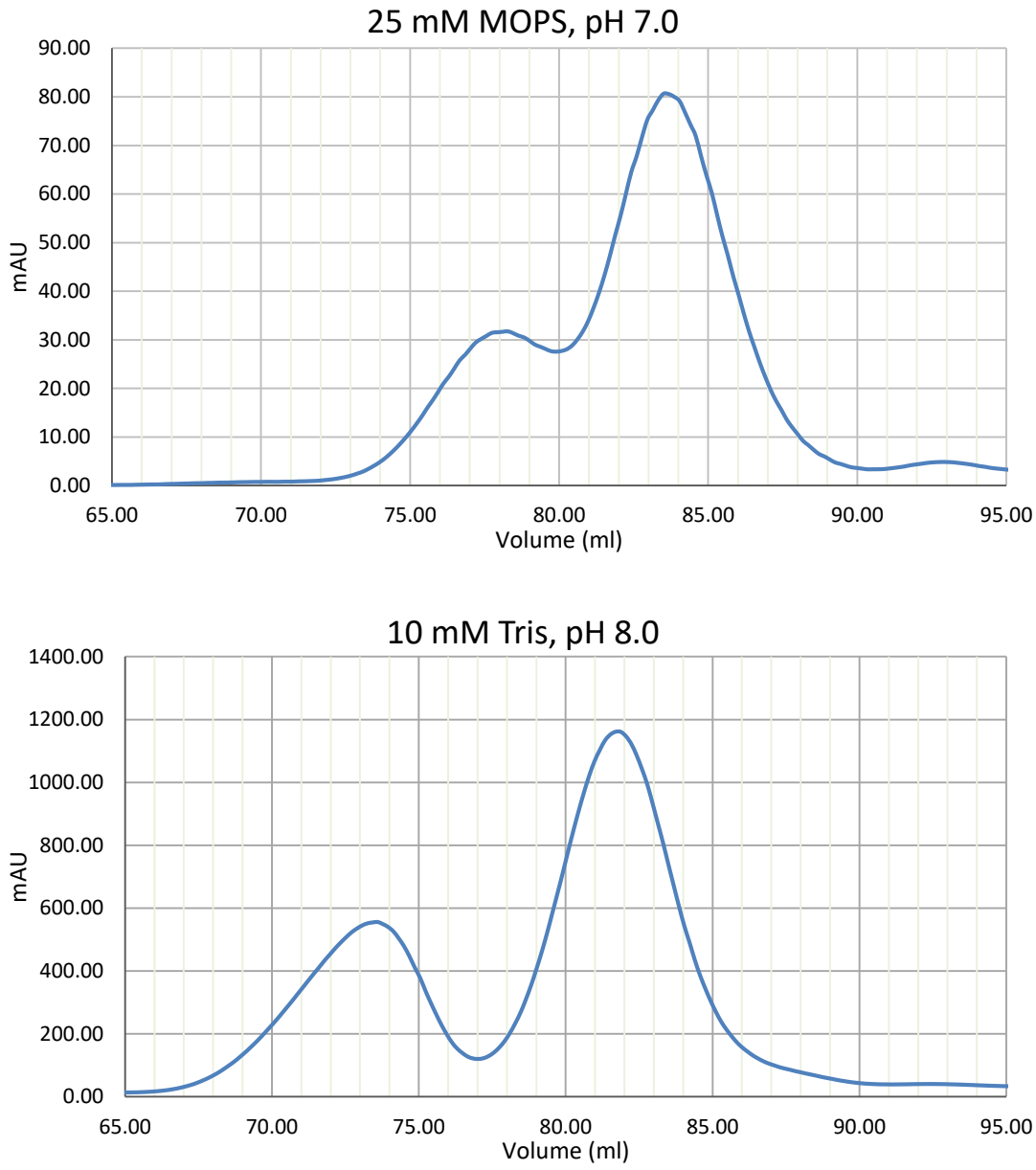


Figure 4.9 Optimisation of size exclusion of Cwp84$_{92\text{-}497_C116A}$. Products from the tryptic digest produced two overlapping peaks (top). The right peak contained pure Cwp84$_{92\text{-}497_C116A}$, while the left peak contained a significant amount of contaminants. Reducing the ionic strength of the buffer resulted in improved peak separation (bottom) and therefore, increased purity of Cwp84$_{92\text{-}497_C116A}$. Decreasing concentration and volume was observed to have no effect on peak separation.

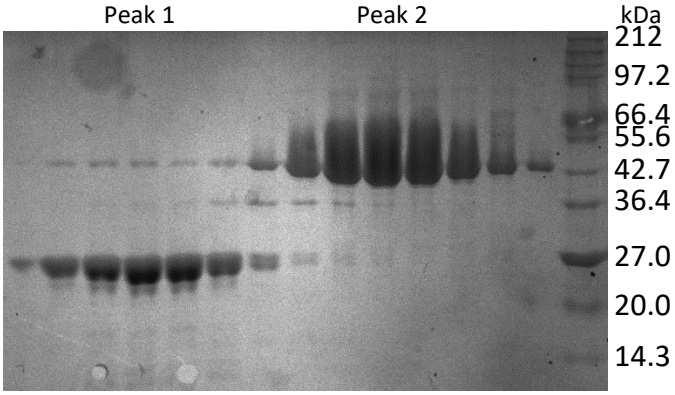


Figure 4.10. SDS-PAGE showing size exclusion chromatography of Cwp84$_{92\text{-}497_C116A}$. Cwp84$_{92\text{-}497_C116A}$ – 45 kDa, GST – 26 kDa. The first peak is GST, while the second peak was confirmed by mass spectrometry to be Cwp84$_{92\text{-}497_C116A}$.

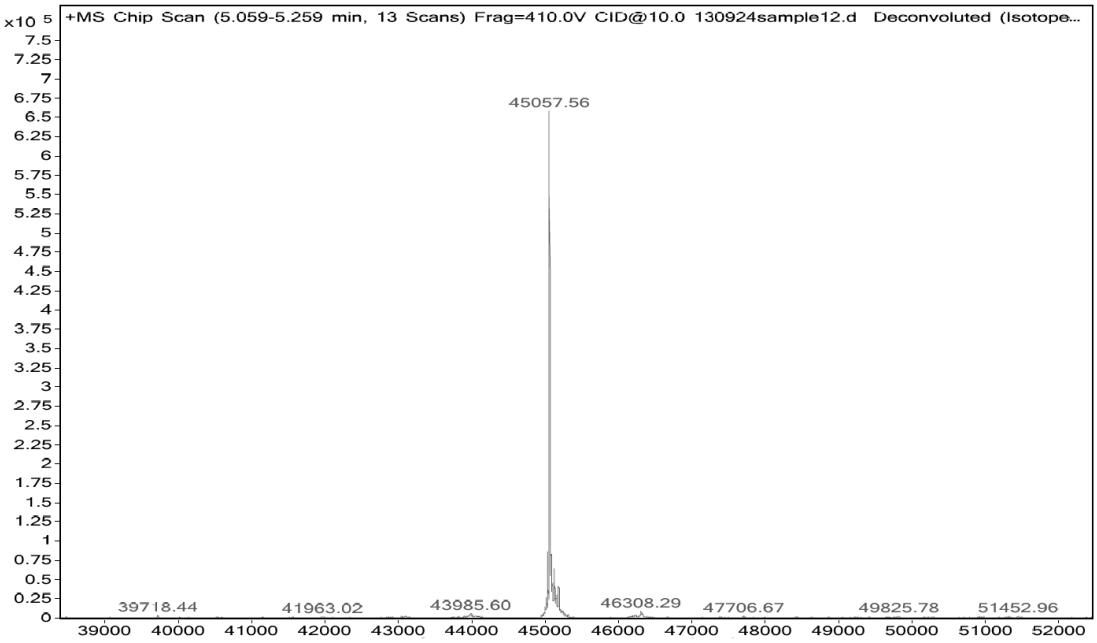


Figure 4.11 Mass spectrum of Cwp84$_{92\text{-}497_C116A}$. The peak at 45 kDa, which corresponds with the predicted mass of the Cwp84 construct minus the 7 kDa propeptide is by far the most abundant species.

4.3.3 Crystallisation

Crystals of construct 1 with the propeptide were obtained in Structure Screen 1&2 condition D7 (0.2 M ammonium sulphate, 30% PEG 4,000). While after propeptide cleavage, the protein crystallised in MIDAS condition F7 (20% dimethyl sulfoxide, 20% Jeffamine M-2070) and PACT condition D7 (0.2 M NaCl, 0.1 M Tris pH 8.0, 20% PEG 6000). The latter was optimised with addition of 10% Silver Bullets condition E9 (0.2% 1,4-diamino-butane, 0.2% cystamine dihydrochloride, 0.2% diloxanide furoate, 0.2% sarcosine, 0.2% spermine, 20 mM sodium HEPES pH 6.8; Hampton Research) (figure 4.12).

4.3.4 Phasing

As initial attempts at molecular replacement were unsuccessful, a $Cwp84_{33\text{-}497\_C116A}$ selenomethionine derivative was produced. The protein was expressed, purified and crystallised in largely the same way, with differences detailed in Chapter 3. Incorporation of selenomethionine was confirmed by electro-spray ionisation mass spectrometry (figure 4.13A).

To confirm the presence of an anomalous signal at the Se-K edge, fluorescence scans were performed at Diamond Light Source, the result from one of these scans is given in figure 4.13B in the form of a *CHOOCH* output (Evans & Pettifer, 2001). X-ray diffraction data were collected at peak, inflection, high remote, and low remote energies. Crystallographic statistics for the selenomethionine datasets are given in table 4.14.

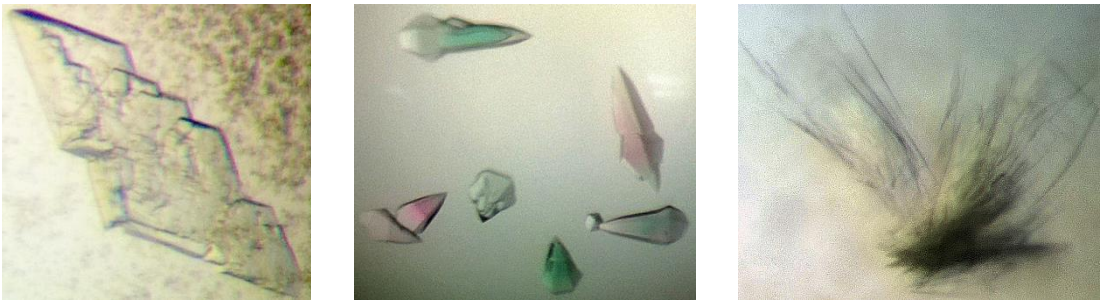
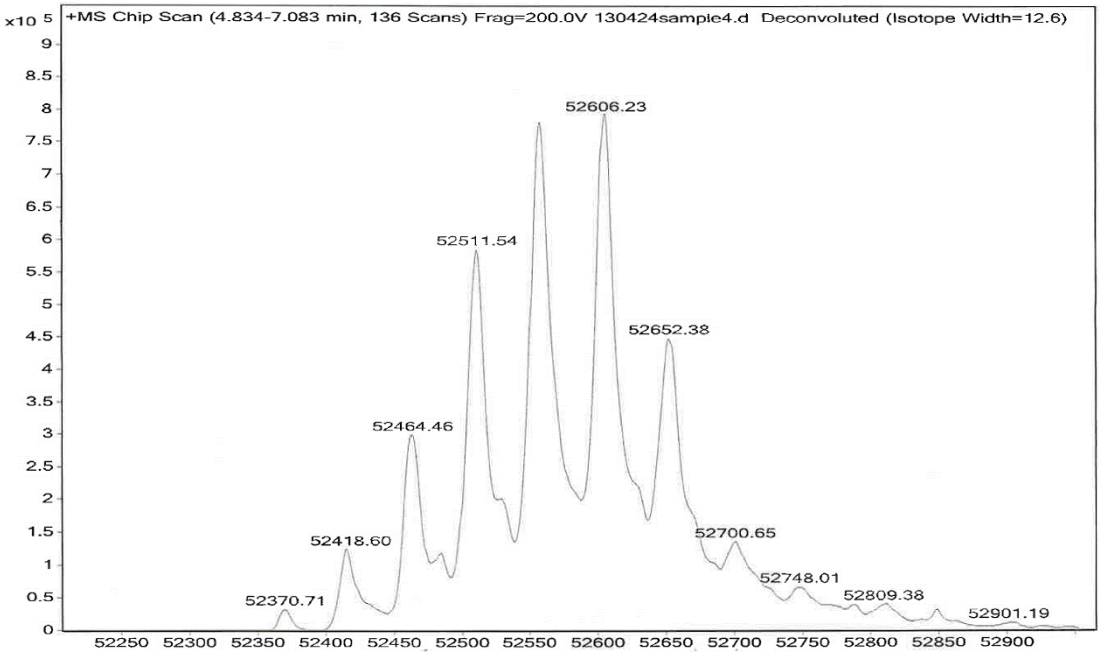


Figure 4.12 $Cwp84_{33\text{-}497\_C116A}$ crystals. Left to right – with the propeptide in Structure screen 1&2 D7, without the propeptide in MIDAS F6, without the propeptide in PACT D7. All pictured crystals were obtained in initial screens.

A



B

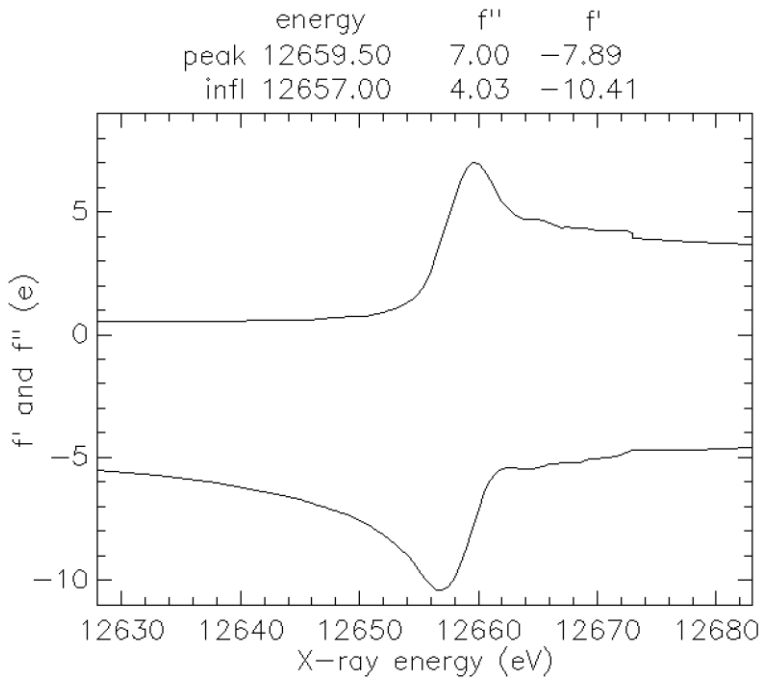


Figure 4.13 Confirmation of Selenomethionine incorporation. (A) Mass spectrum. The peak at 52371 corresponds to $Cwp84_{33\text{-}497\_C116A}$ with 6 selenomethionine residues and 6 sulphomethionine residues, while 52652 corresponds to 12 selenomethionines. It can be estimated from this that the average $Cwp84_{33\text{-}497\_C116A}$ molecule contained approximately 10 selenomethionine residues. **(B)** Fluorescence scan *CHOOCH* output. This gives the peak and inflection energies at which anomalous data should be collected and the respective anomalous scattering factors that can be used in structure determination.

Table 4.14A Crystallographic statistics for selenomethionine datasets. Inner shell statistics are given in square brackets. Outer shell statistics are given in round brackets.

	Peak	**Inflection**
Energy (eV)	12,660	12,656
Wavelength (Å)	0.9793	0.9796
Space group	$P2_1$	$P2_1$
Unit-cell parameters		
a (Å)	50.9	51.0
b (Å)	73.1	73.2
c (Å)	125.4	125.7
α = γ (°)	90.0	90.0
β (°)	93.5	93.5
Resolution range (Å)	[29.66 - 6.64] (2.21 - 2.10)	[29.66 - 6.64] (2.21 - 2.10)
R_{merge}	[0.193] 0.322 (0.575)	[0.169) 0.325 (0.576)
R_{meas}	[0.203] 0.338 (0.600)	[0.181] 0.343 (0.622)
R_{pim}	[0.047] 0.104 (0.134)	[0.040] 0.109 (0.233)
I/σ(I)	[34.6] 16.9 (7.7)	[31.9] 16.0 (6.3)
Completeness (%)	[98.9] 100.0 (100.0)	[98.9] 100.0 (100.0)
Number of reflections	[34,390] 1,120,802 (160,996)	[35,441] 945,054 (110,376)
Unique reflections	[1,758] 52,790 (7,864)	[1,766] 54,120 (7,881)
Multiplicity	[19.6] 20.8 (20.5)	[20.1] 17.5 (14.0)
Anomalous completeness (%)	[99.2] 100.0 (100.0)	[99.2] 100.0 (100.0)
Anomalous multiplicity	[10.3] 10.5 (10.2)	[10.5] 8.8 (6.9)
CC_{anom}	[0.512] 0.218 (0.077)	[0.569] 0.174 (0.060)
CC_{anom} < 0.3 (Å)	3.8	4.4

Table 4.14B Crystallographic statistics for selenomethionine datasets. Inner shell statistics are given in square brackets. Outer shell statistics are given in round brackets.

	High remote	Low remote
Energy (eV)	12,770	12,550
Wavelength (Å)	0.9717	0.9879
Space group	$P2_1$	$P2_1$
Unit-cell parameters		
a (Å)	50.7	51.3
b (Å)	73.0	73.6
c (Å)	125.7	125.4
α = γ (°)	90.0	90.0
β (°)	93.9	93.1
Resolution range (Å)	[29.60 - 7.91] (2.64 - 2.50)	[29.27 - 6.15] (2.05 – 1.94)
R_{merge}	[0.091] 0.185 (0. 542)	[0.052] 0.139 (0.815)
R_{meas}	[0.101] 0.205 (0.603)	[0.060] 0.162 (0.976)
R_{pim}	[0.043] 0.063 (0.188)	[0.022] 0.083 (0.525)
I/σ(I)	[28.5] 14.5 (5.1)	[21.8] 9.2 (2.0)
Completeness (%)	[98.2] 99.9 (100.0)	[99.1] 99.4 (96.4)
Number of reflections	[10,391] 333,693 (47,448)	[16,240] 489,672 (53,896)
Unique reflections	[1,048] 31,917 (4,644)	[2,260] 68,579 (9,692)
Multiplicity	[9.9] 10.5 (10.2)	[7.2] 7.1 (5.6)
Anomalous completeness (%)	[98.5] 99.9 (100.0)	[98.7] 98.2 (90.7)
Anomalous multiplicity	[5.3] 5.3 (5.1)	[3.7] 3.6 (2.9)
CC_{anom}	[0.453] 0.112 (0.106)	[-0.204] -0.076 (-0.023)
CC_{anom} < 0.3 (Å)	5.5	N/A

CRUNCH2 found 55 potential selenium sites out of a predicted 48 within the unit cell, this resulted in the determination of initial phases, from which model building and refinement could proceed, allowing *Buccaneer* and REFMAC5 to produce an output model with a figure of merit of 0.856 and R_{work} and R_{free} values of 0.248 and 0.277, respectively. The resulting structure was refined against the higher resolution native data, for which statistics are given in Table 4.15.

Table 4.15 Crystallographic statistics for Cwp84$_{33\text{-}497_C116A}$. Inner shell statistics are given in square brackets. Outer shell statistics are given in round brackets.

Space group	**$P2_1$**
Unit-cell parameters	
a (Å)	50.9
b (Å)	73.5
c (Å)	125.6
α = γ (°)	90.0
β (°)	93.6
Resolution range (Å)	[48.15 - 7.67] (1.42 - 1.40)
R_{merge}	[0.138] 0.099 (0.258)
R_{meas}	[0.171] 0.121 (0.349)
R_{pim}	[0.058] 0.054 (0.301)
$CC_{1/2}$	[0.936] 0.989 (0.679)
I/σ(I)	[26.5] 16.0 (4.2)
Completeness (%)	[97.9] 93.9 (65.2)
Number of reflections	[5,982] 810,986 (14,198)
Unique reflections	[1,146] 170,213 (5,848)
Multiplicity	[5.2] 4.8 (2.4)
R_{work}/R_{free}	0.138/0.169
RMSDs	
Bond Lengths (Å)	0.008
Bond Angles (°)	1.340
Ramachandran Statistics	
Favoured	96.1
Allowed	3.9
Outliers	0
Average B-factors (Å^2)	
Protein	16.7
Ligand	36.6
Water	29.8
Number of Atoms	
Protein	7404
Ligand	104
Water	928
PDB code	4CI7

4.3.5 Structure with the propeptide

The structure of a truncated (residues 33-497) Cwp84 active site mutant (C116A) has been solved to a resolution of 1.4 Å (Bradshaw *et al.*, 2014). The structure consists of the propeptide, the cysteine protease domain and a newly identified "lectin-like" domain, named for its structural similarity to the superfamily of carbohydrate binding proteins (figure 4.16).

The structure was solved in the monoclinic space group $P2_1$ with two molecules in the asymmetric unit. It also contains two calcium ions, two sulphate ions, eight PEG molecules, six glycerol molecules and 927 water molecules, with a solvent content of 43.8%. The structure was refined to R_{work} and R_{free} values of 0.138 and 0.169, respectively. Poor electron density was observed between Gly58 and Tyr63 although this area could be modelled with a fair degree of certainty. Little to no density was observed for Lys81 to Tyr89, so this region was not modelled. The two chains superpose on each other with an RMSD of 0.19 Å (2910 atoms).

4.3.6 Structures without the propeptide

A model based on the structure with the propeptide was used for molecular replacement to determine the structure of $Cwp84_{92\text{-}497\_C116A}$ using data from two crystal forms (Bradshaw *et al.*, 2015). Crystallographic statistics for these data are summarised in table 4.17.

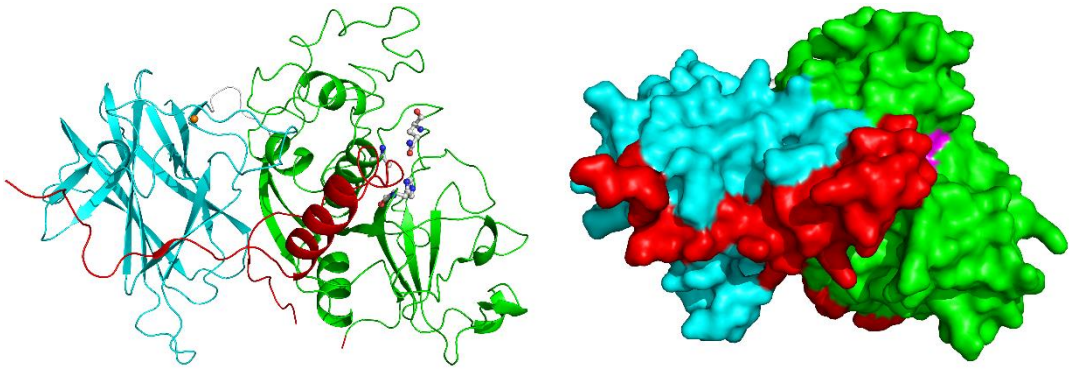


Figure 4.16 The structure of $Cwp84_{33\text{-}497\_C116A}$. The cysteine protease domain is coloured in green with the active site Q110, C116A and H262 shown as sticks on the left and in pink on the right. The lectin-like domain is shown in cyan with the calcium ion in orange and the propeptide is coloured red. The cysteine protease domain assumes a cathepsin L-like fold.

Table 4.17 Crystallographic statistics for Cwp84$_{92\text{-}497_C116A}$. Inner shell statistics are given in square brackets. Outer shell statistics are given in round brackets.

	Structure 1	Structure 2
Space group	P1	P1
Unit-cell parameters		
a (Å)	42.2	48.1
b (Å)	58.4	70.2
c (Å)	93.1	78.9
α (°)	89.3	65.2
β (°)	78.0	89.9
γ (°)	71.6	80.2
Resolution range (Å)	[55.35 – [9.01] (1.88 - 1.84)	[47.3 - 8.76] (1.63 - 1.60)
R_{merge}	[0.024] 0.071 (0.589)	[0.058] 0.092 (0.467)
R_{meas}	[0.033] 0.101 (0.833)	[0.082] 0.131 (0.661)
R_{pim}	[0.024] 0.071 (0.589)	[0.058] 0.092 (0.467)
$CC_{1/2}$	[0.994] 0.996 (0.610)	[0.993] 0.991 (0.655)
I/σ(I)	[40.3] 13.6 (2.7)	[11.8] 6.8 (2.1)
Completeness (%)	[82.3] 83.5 (83.0)	[96.1] 90.8 (49.9)
Number of reflections	[1,525] 155,876 (9,319)	[2,125] 313,196 (7,586)
Unique reflections	[508] 59,711 (3,659)	[709] 110,175 (2,994)
Multiplicity	[3.0] 2.6 (2.5)	[3.0] 2.8 (2.5)
R_{work}/R_{free}	0.227/0.291	0.181/0.210
RMSDs		
Bond Lengths (Å)	0.009	0.009
Bond Angles (°)	1.323	1.209
Ramachandran Statistics (%)		
Favoured	96.5	95.4
Allowed	3.5	4.6
Outliers	0	0
Average B-factors (Å^2)		
Protein	31.9	14.7
Ligand	31.6	30.3
Water	32.0	26.5
Number of Atoms		
Protein	6394	6662
Ligand	18	77
Water	449	791
PDB code	4D59	4D5A

The first of the two structures was solved at 1.84 Å with two protein chains in the asymmetric unit, two calcium ions, two jeffamine molecules, and 449 water molecules. The second was solved at 1.6 Å and contained two protein chains, four calcium ions, eight PEG molecules, four glycerol molecules and 791 water molecules. The four chains superpose on each other with RMSDs between 0.12 Å (2619 atoms) and 0.47 Å (2647 atoms).

Apart from the obvious lack of a propeptide, the two structures without the propeptide are largely unchanged when compared to that with the propeptide (figure 4.18), superposing with RMSDs between 0.40 Å (2542 atoms) and 0.62 Å (2585 atoms). There are however, three loops that undergo notable conformational changes: Met160-Ser164 and Leu315-Asn321 in both structures, which both form part of the active site groove and Thr479-Pro485, which is presented on the surface of the lectin-like domain, in both structures but to a much greater extent in the second structure.

4.3.7 Propeptide structure

SignalP identifies the signal peptide cleavage site of Cwp84 as being between Ala30 and Glu31 (Petersen *et al.*, 2011), while the propeptide cleave site has been shown to be between Lys91 and Ser92 (de la Riva *et al.*, 2011; Bradshaw *et al.*, 2014). The structure presented here has all residues of the propeptide from His33, however Asn80 to Thr90 are not visible. The C-terminal portion of the propeptide (Val66 to Lys91) assumes a typical cysteine protease propeptide fold (Coulombe *et al.*, 1996; Sivaraman *et al.*, 1999), with an extended loop conformation running down the active site groove in the opposite direction to that of a substrate (figure 4.19). The N-terminal portion (His33 to Gly65), on the other hand, assumes a novel conformation, folding back on itself and wrapping around the lectin-like domain, instead of interacting with the prosegment binding loop (PBL) on the cysteine protease domain and forming a small globular domain, as would normally be seen (figure 4.20) (cf. figure 4.2). This leaves the top of the active site grove considerably more exposed than is seen in other cysteine proteases. Literature and DALI searches did not reveal any previous cysteine protease structures that exhibit this propeptide conformation.

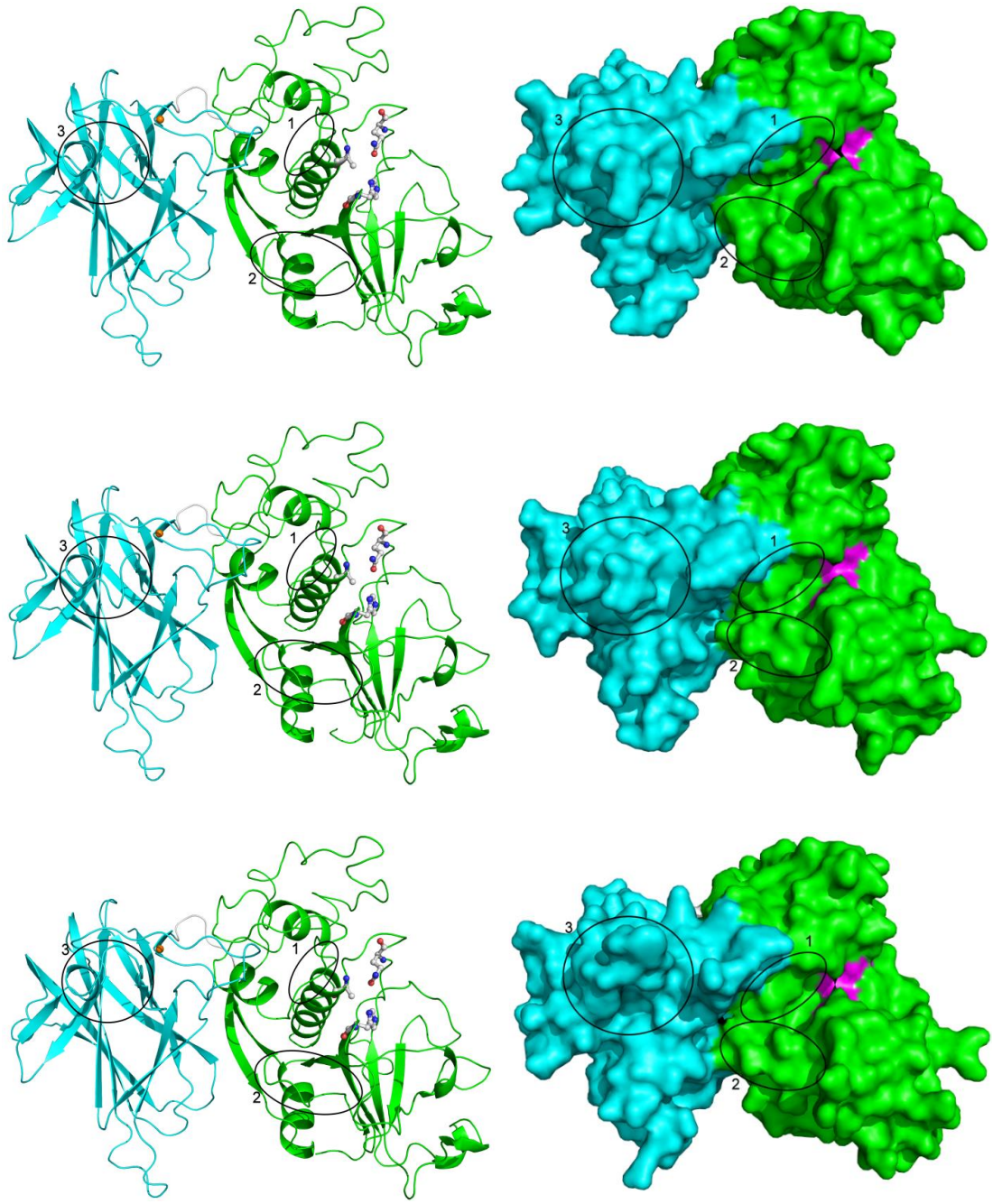


Figure 4.18 Structures of Cwp84 construct 1 without the propeptide. The structures are represented with a ribbon diagram and a surface. The cysteine protease domain is coloured in green with the active site Q110, C116A and H262 shown as sticks on the left and in pink on the right. The three loops that undergo conformational changes upon propeptide cleavage are circled. 1: Met160-Ser164, 2: Leu315-Asp320, 3: Thr479-Pro485. **(A)** The structure with the propeptide graphically removed. **(B)** The first structure without the propeptide **(C)** The second structure without the propeptide. Both possess differing conformations in loops 1 and 2, while C also shows a significantly different conformation in loop 3.

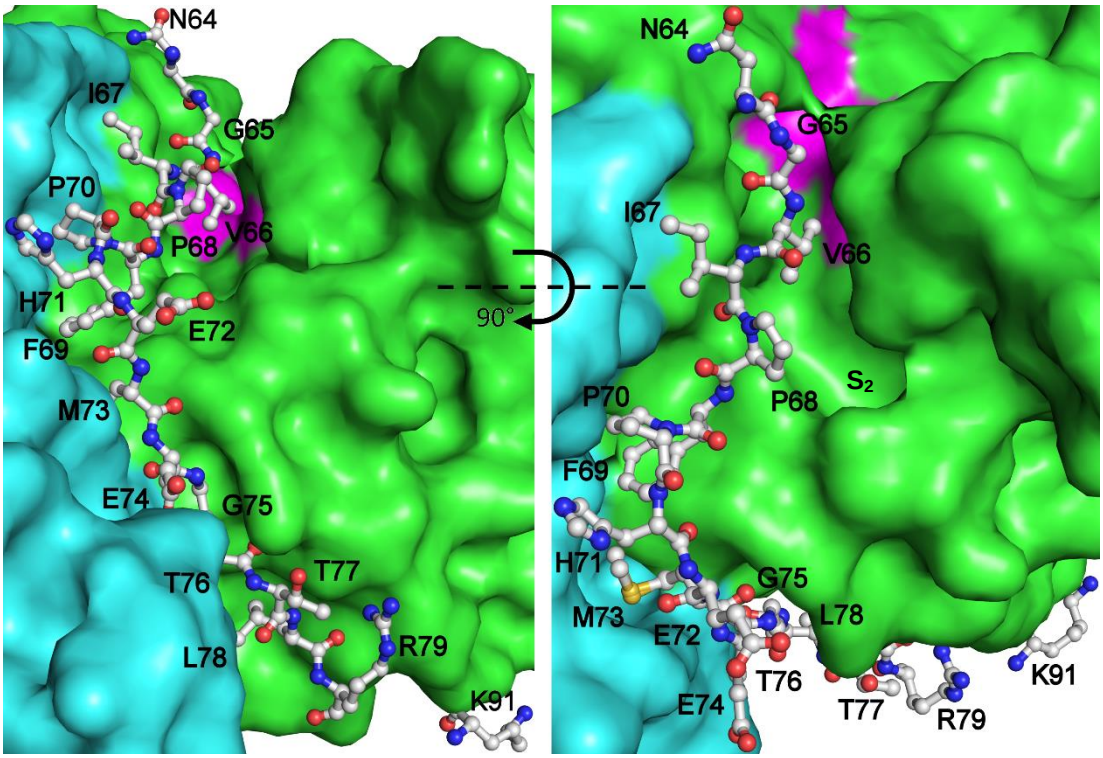


Figure 4.19 The active site groove of Cwp84. The cysteine protease domain is coloured in green with the lectin-like domain in cyan. The S_2 pocket, which is involved in substrate recognition is indicated. The C-terminal portion of the propeptide, represented as sticks, sits in the active site groove, part of which is formed by the lectin-like domain, resulting in a somewhat deeper groove than is normally seen in cysteine proteases. The effect that the presence of the lectin-like domain has on catalysis is currently unknown.

4.3.8 Cysteine Protease domain structure

The cysteine protease domain of $Cwp84_{33\text{-}497\_C116A}$ assumes a typical two lobed fold, with one half formed primarily by a β-sheet and the other primarily by a mixture of loops and α-helices. The active site groove is found between the two lobes. A DALI search showed that the highest level of similarity is to cathepsin L-like cysteine proteases. There are, however, some notable differences, namely, the aforementioned lack of a PBL, a slightly different occluding loop structure, and a deeper active site cleft, partially caused by the presence of the lectin-like domain.

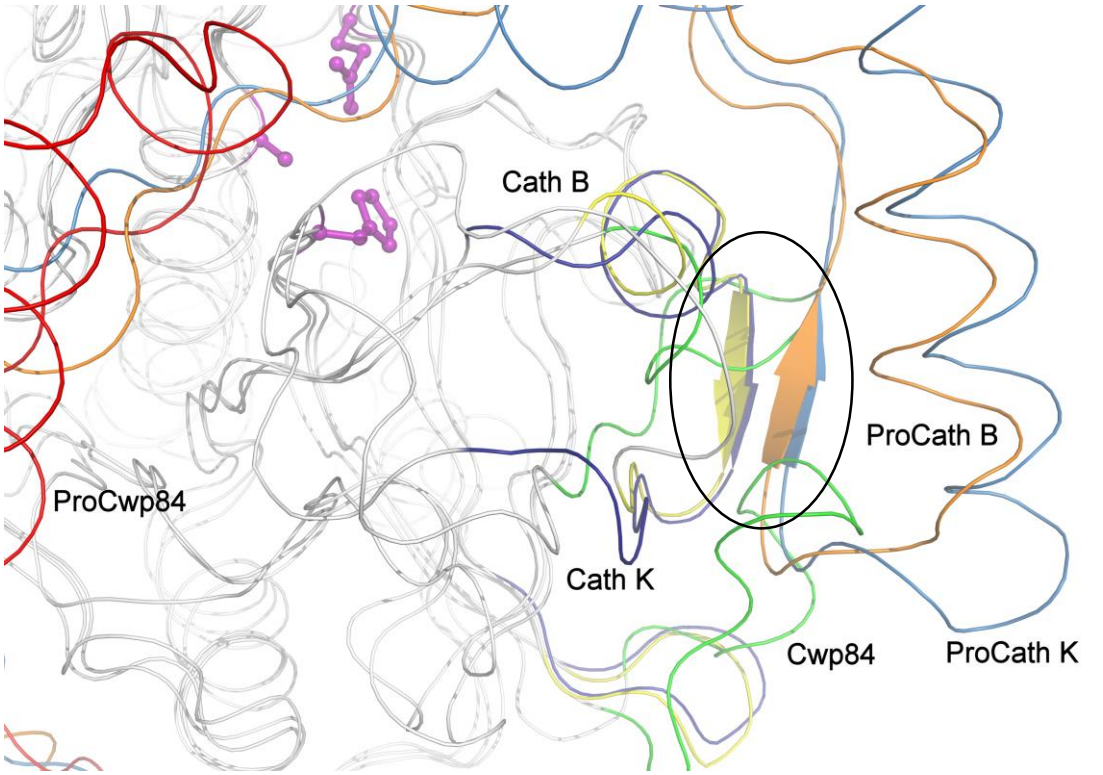


Figure 4.20 Comparison of prosegment binding loops (PBLs). The PBL region is circled. The PBL of cathepsin K (cathepsin L-like, 7PCK) is shown in dark blue with the propeptide in light blue. The PBL of cathepsin B (1PBH) is shown in yellow with the propeptide in orange. The equivalent region of Cwp84 is shown in green with the propeptide in red. The active site residues of Cwp84 are shown in magenta. The PBL folds in cathepsin K and cathepsin B are largely the same, as are the propeptide folds in this portion, allowing the propeptide to be anchored to the PBL. The propeptide of Cwp84, which sits in the active site groove, like that of other cysteine proteases, does not wrap around the cysteine proteases domain, which results in a lack of a need for a PBL. As a result, this region assumes a different fold, with a short loop that normally sits below the PBL becoming considerably extended.

4.3.9 Lectin-like domain structure

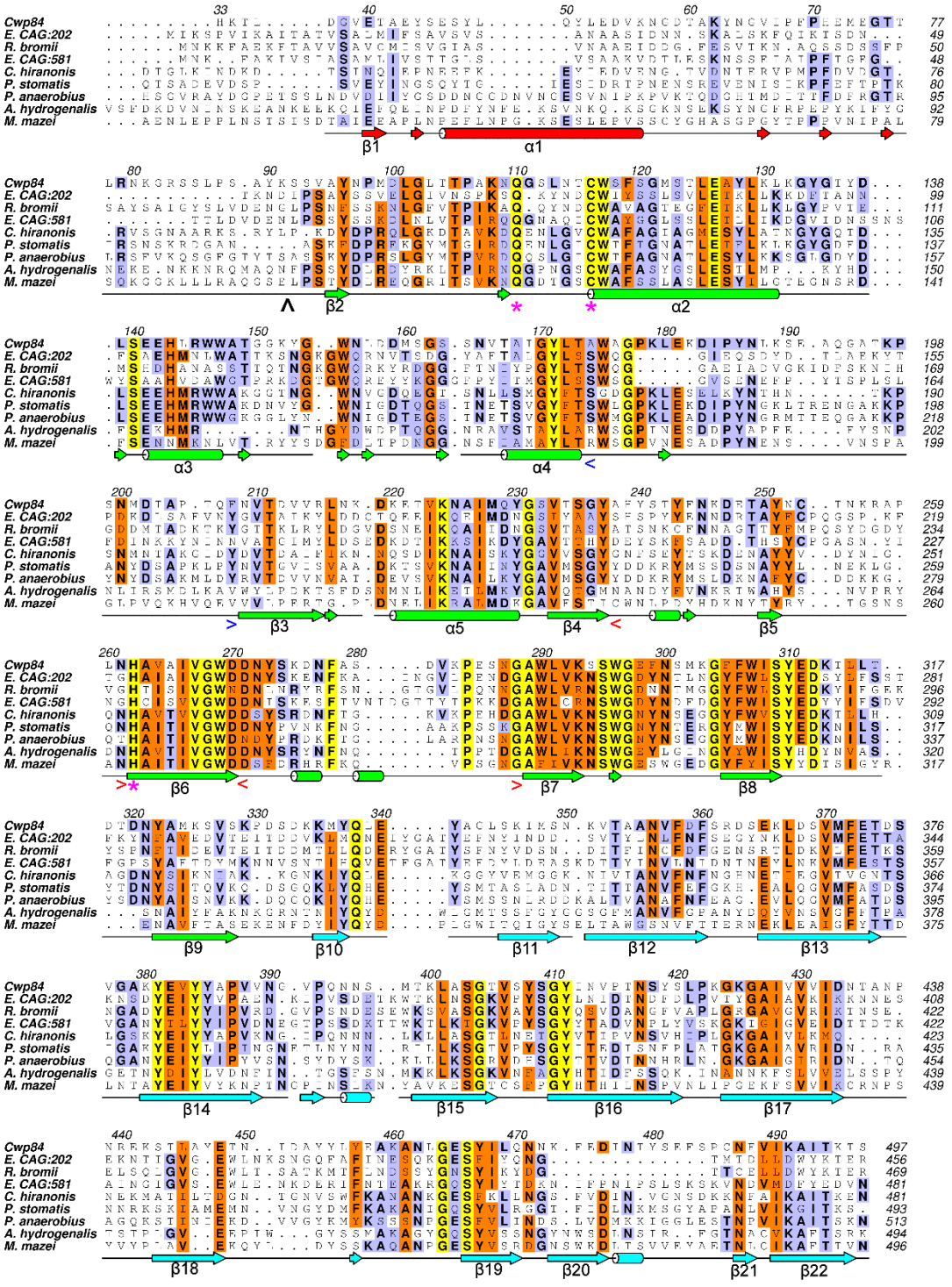
The region of approximately 170 residues with a previously unknown structure found between the cysteine protease domain and the first cell wall binding domain assumes a twisted β-sandwich fold that is stabilised by a calcium ion. This newly identified domain bears some similarity to carbohydrate binding proteins (many of the highest results on a DALI search were carbohydrate binding proteins), so it has been dubbed the "lectin-like" domain. A BLAST search against the lectin-like domain revealed a range of other predicted cysteine proteases that also appear to possess a lectin-like domain. The combination of a cysteine protease domain and lectin-like domain

appears to be present in a range of species across the *Clostridiales* order and is also seen in some archaea (figure 4.21). The lectin-like domain contains a calcium ion coordinated by Leu339, Glu448, Lys460, Asn487 and two water molecules. The majority of conserved residues are found within β-sheets, so are likely to be involved in the generation of tertiary structure, while loop regions are considerably more variable.

4.3.10 Co-crystallisation

To further understand substrate and inhibitor binding, screens were set up after soaking $Cwp84_{33\text{-}497\_C116}$ with E-64 and with two peptides based on the cleavage site of SlpA. The peptides used the sequence from lab strain 630 – LETKS|ANDTI and the sequence from the hyper virulent ribotype 027 – RUTTKS|AAKASI (vertical lines indicate the cleavage site). Crystals grew and useable data were collected but no significant density was observed in the active site for any dataset.

Figure 4.21 (Next page) Multiple sequence alignment of $Cwp84_{33–497}$ and the highest BLAST results. All are cysteine proteases that possess a putative lectin-like domain. The alignment was performed using ClustalW2 (Larkin *et al.*, 2007) and rendered with ALINE (Bond & Schuttelkopf, 2009). Strictly conserved residues are shown in yellow, medium to well conserved residues are in orange and slightly conserved residues are in blue. The secondary structure of Cwp84, as predicted by DSSP (Kabsch & Sander, 1983) is also shown with the propeptide in red, the cysteine protease domain in green and the lectin-like domain in blue. 3_{10} helices and β-bridges are displayed in the same way as α-helices and β-strands, but are not numbered. Active-site residues (Gln110, Cys116 and His262) are indicated with pink stars, the propeptide cleavage site (Lys91-Ser92) is indicated with a black arrow and the occluding loop and PBL regions are indicated with blue and red triangular brackets, respectively. Sequences are taken from the following NCBI GenBank references: Cwp84, NC_009089; *Eubacterium CAG:202*, CDC03302; *Ruminococcus bromii*, YP_007780613; *Eubacterium CAG:581*, CDF12829; *Clostridium hiranonis*, WP_006441026; *Peptostreptococcus stomatis*, WP_007788460; *P. anaerobius*, WP_002842957; *Anaerococcus hydrogenalis*, WP_004816163; *Methanosarcina mazei*, NP_632235. The proteins from *C. hiranonis*, *P. stomatis* and *P. anaerobius* possess three putative Pfam 04122 repeats and thus are likely to be S-layer proteins performing similar functions to Cwp84.

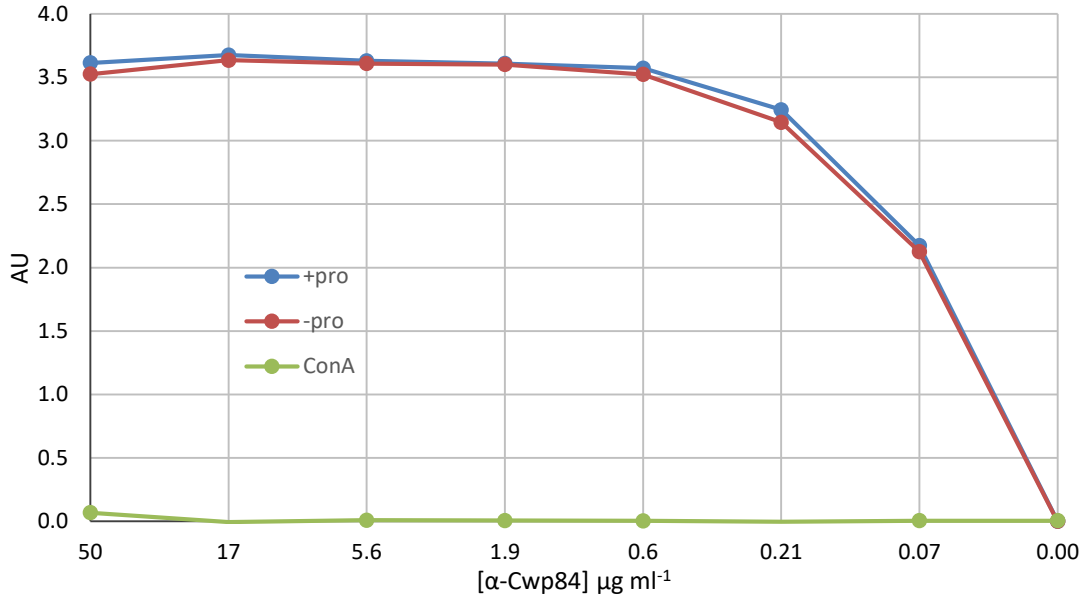

Cwp84
E. CAG:202
R. bromii
E. CAG:581
C. hiranonis
P. stomatis
P. anaerobius
A. hydrogenalis
M. mazei
β1
α1
β2
α2
α3
α4
β3
α5
β4
β5
β6
β7
β8
β9
β10
β11
β12
β13
β14
β15
β16
β17
β18
β19
β20
β21
β22

4.3.11 Lectin-like domain ELISA

An ELISA was performed to determine the ability of the lectin-like domain of Cwp84 to bind to a range of sugars. Antibodies were raised against $Cwp84_{92\text{-}497\_C116A}$, while concanavalin A, which is known to selectively bind to glucose and mannose, was used as a control. A preliminary ELISA was performed to check binding of antibodies against $Cwp84_{92\text{-}497\_C116A}$ and concanavalin A to their respective targets, to determine an appropriate concentration for the main ELISA and to confirm that there was no cross-reactivity. This showed appropriate concentrations of 0.6 μg ml^{-1} for α-$Cwp84_{92\text{-}497\_C116A}$ and 2.0 μg ml^{-1} for α-concanavalin A. It also confirmed specific binding of the antibodies (figure 4.22). The two proteins were then screened for their binding to specific carbohydrates. A range of carbohydrates were bound to a 96 well block, this was washed with the two proteins and antibodies against them. Theoretically, if Cwp84 was able to bind any of the carbohydrates tested, these wells would be highlighted by the ELISA. A negative control was performed without the addition of any proteins. A difference in signal was seen between the two proteins and for the negative control, indicating that the presence of a protein and the identity of the protein made a difference, however, no stronger signal was seen for any sugar for $Cwp84_{33\text{-}497\_C116A}$, $Cwp84_{92\text{-}497\_C116A}$ or for concanavalin A, indicating that the protocol did not work correctly (figure 4.23).

4.3.12 Expression of constructs 2 to 4

Constructs coding for the cell wall binding domains have previously been observed to result in problematic expression and purification. This is why construct 1 was originally created. However, it was decided after the structure of the portion of Cwp84 coded for by construct 1 had been determined, that work should be performed on the other constructs in an attempt to overcome some of these problems so that the cell wall binding domains could be studied.



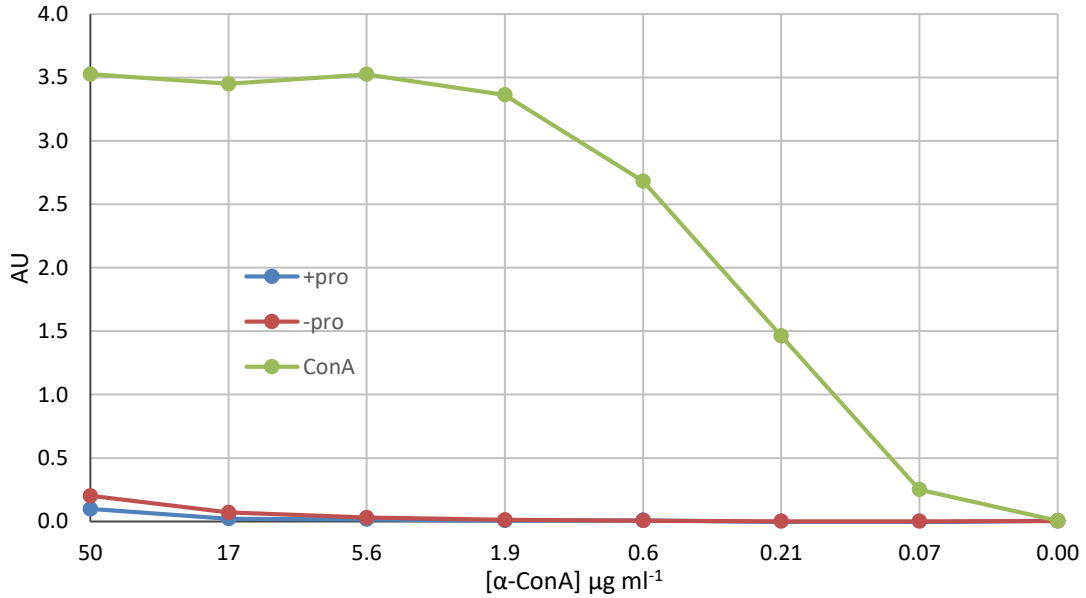


Figure 4.22 Determination of optimal IgG concentration. Antibodies against $Cwp84_{92\text{-}497\_C116A}$ and concanavalin A were screened against both proteins ($Cwp84_{33\text{-}497\_C116A}$ both with (+) and without (-) the propeptide) to determine an appropriate concentration and to confirm specific binding. The antibodies can be seen to only bind to their target protein. α $Cwp84_{92\text{-}497\_C116A}$ was used at 0.6 µg ml^{-1} αConA was used at 2.0 µg ml^{-1}.

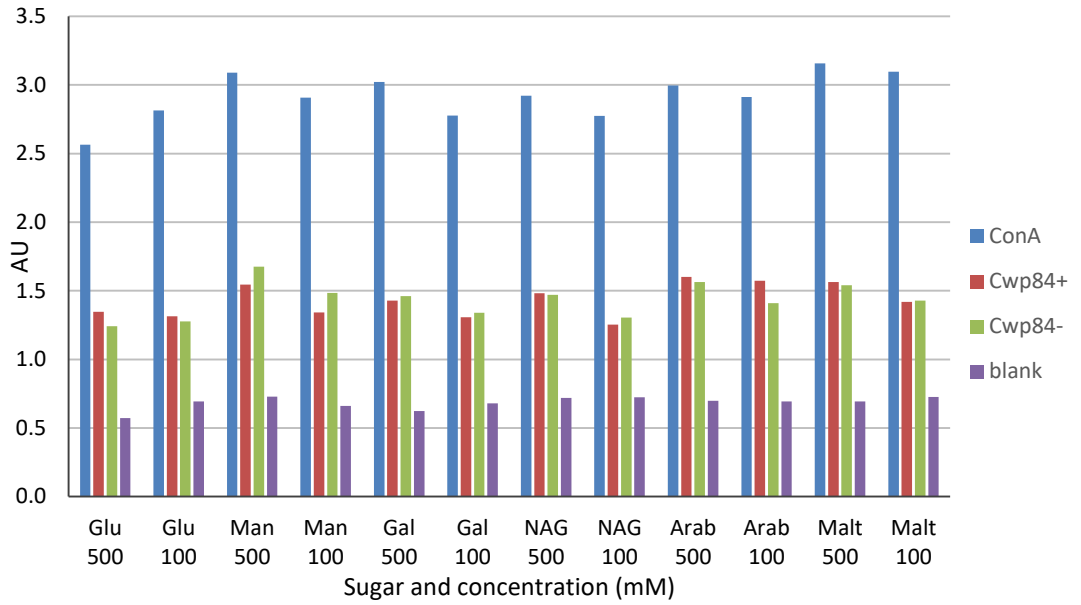


Figure 4.23 Carbohydrate binding assay. Concanavalin A and Cwp84$_{33\text{-}497_C116A}$, both with (+) and without (-) the propeptide, were screened for their ability to bind to a range of sugars at 500 mM and 100 mM. Glu – glucose. Man – manose. Gal – galactose. NAG – N-acetylglucosamine. Arab – arabinose. Malt – Maltose. None showed any difference in binding including the positive controls of concanavalin A with glucose and mannose.

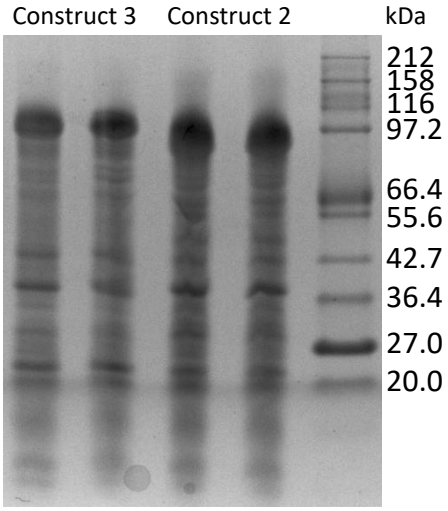


Figure 4.24 SDS PAGE showing expression of construct 2 and 3. Two samples are shown for each of the constructs just before harvesting. Construct 3 was predicted to be 100 kDa, while construct 2 was predicted to be 89 kDa. Large bands are visible at the appropriate sizes, indicating that a large amount of protein was expressed without degradation. Construct 4, coding for full length Cwp84$_{33\text{-}803_C116A}$ with a GST tag appeared similar and was appropriately larger.

Initial expression trials for constructs 2 to 4 were performed at 16 °C overnight and resulted in a significant degree of degradation. It was found that expressing for four hours at 37° C resulted in insoluble expression, which lead to a significant reduction in the degree of degradation (figure 4.24), however it necessitated the development of a refolding protocol.

4.3.13 Purification of constructs 2 to 4

Initial attempts at purifying a longer construct used construct 4, coding for full length Cwp84. This began with a standard inclusion body washing protocol to remove soluble and membrane components followed by solubilisation in 8 M urea. As the protein was GST tagged, it was refolded by a two-step dialysis protocol into GST loading buffer. Analysis by SDS-PAGE indicated a good amount of protein in samples containing inclusion bodies and solubilised protein but not in samples containing refolded protein. The refolding process was slowed by increasing the amount of steps and more gradually reducing the urea concentration. As the structures from construct 1 had shown that Cwp84 binds a calcium ion, $CaCl_2$ was also added to some buffers in the refolding process. This resulted in the development of the refolding protocol given in Chapter 3. Ultimately, this yielded a much greater amount of refolded protein.

The refolded full length protein could be purified to a point that it was by far the most abundant species with the normal GST purification protocol (figure 4.25). When attempts were made to complete the purification with size exclusion chromatography on a GE Healthcare Superdex 16/60, all of the full length protein from construct 4 was observed to elute after approximately 46 ml, the void volume of the column, indicating that the protein was suffering from severe aggregation. Attempts were made to adjust the buffer including addition of a different chaotropes, kosmotropes, detergents, reducing agents, buffers and sugars but the aggregation persisted. It was decided that constructs 2 and 3, with fewer cell wall binding domains should be purified by the same protocol as they may not have suffered from the same aggregation problems. These constructs could be purified to a similar degree to construct 4 and, although the majority of the purified protein was still observed to

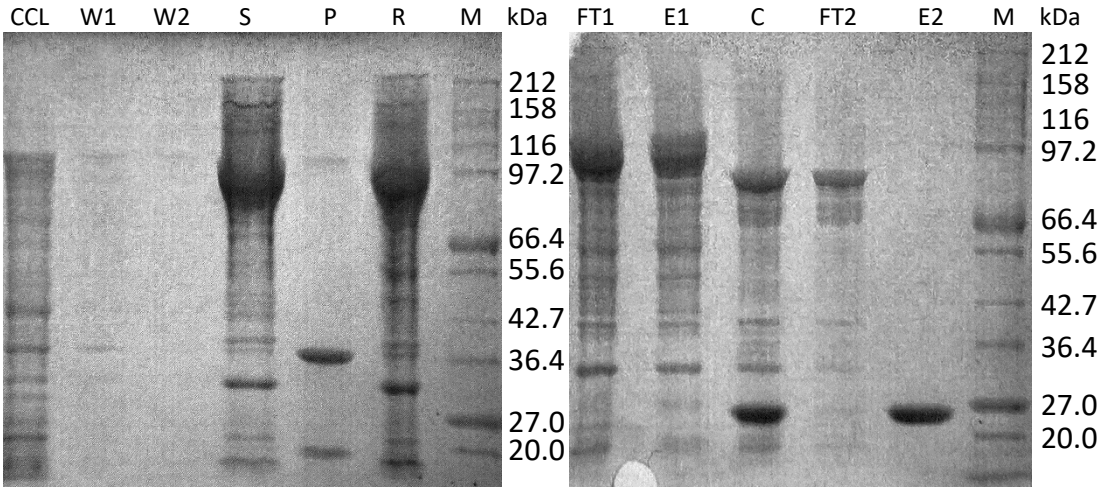


Figure 4.25 SDS-PAGE showing purification of construct 4. $Cwp84_{33\text{-}803\_C116A}$ with GST tag – 110 kDa, $Cwp84_{33\text{-}803\_C116A}$ without GST tag – 84 kDa. CCL – Cleared cell lysate. W1 – supernatant after 1st inclusion body wash. W2 – Supernatant after 2nd inclusion body wash. S – Supernatant after solubilisation. P – Pellet after solubilisation. R – Refolded protein after serial dialysis. FT1 – Flow through from 1st round of GST purification. E1 – Eluate from 1st round of GST purification. C – Cleaved protein after dialysis with 3C protease. FT2 – Flow through from 2nd round of GST purification. E2 – Eluate from 2nd round of GST purification. M – Markers.

form aggregates during SEC, some protein did elute after the correct volume (figure 4.26). This was shown to be pure protein (figure 4.27), which was taken forward for crystallisation and Small angle X-ray scattering (SAXS) studies.

4.3.14 Structural analysis of constructs 2 to 4

As the full length construct could not be purified without aggregating, construct 3 was the primary construct of interest. Crystallisation screens were set up with the protein and did yield multiple crystals that diffracted well, however, when the structure was solved using these data, no density was observed for the cell wall binding domains indicating that they had degraded, resulting, effectively, in the same protein as construct 1.

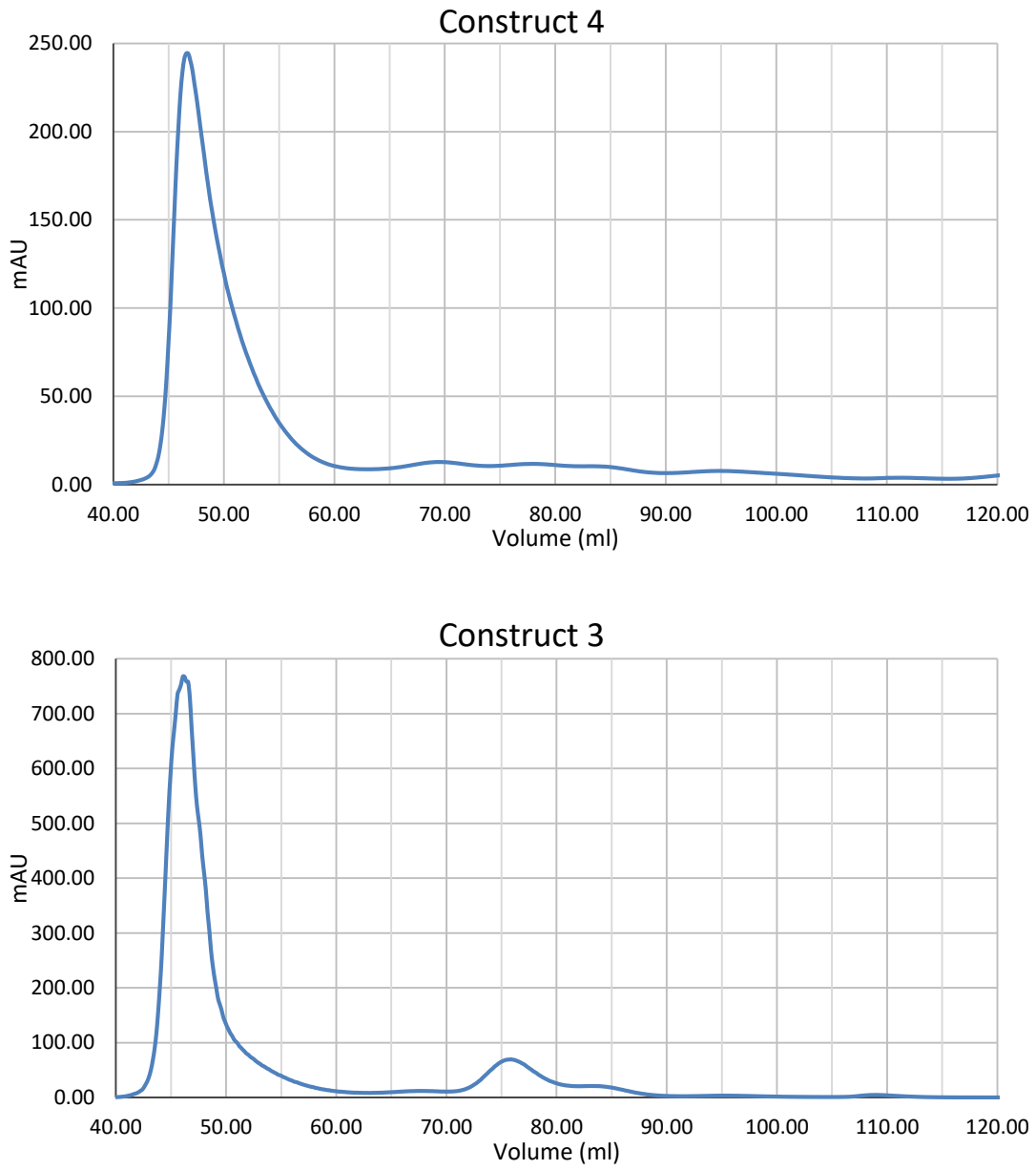


Figure 4.26 Chromatograms for size exclusion of constructs 3 and 4. Both constructs produced a peak after approximately 46 ml, the void volume of the column, indicating aggregation. Construct 3 ($Cwp84_{32\text{-}700\_C116A}$), however, which lacked the third cell wall binding domain produced a small peak after approximately 76 ml of non-aggregated pure protein. A similar peak was not observed for construct 4, coding for full length Cwp84 ($Cwp84_{32\text{-}803\_C116A}$) but was observed for construct 2 ($Cwp84_{32\text{-}600\_C116A}$), which lacked the second and third cell wall binding domains.

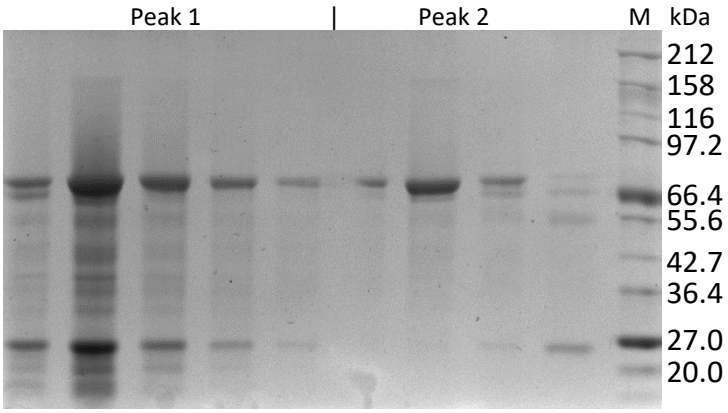


Figure 4.27 SDS-PAGE showing size exclusion of construct 3. $Cwp84_{33\text{-}700\_C116A}$ – 73 kDa. The first peak contains aggregated and degraded protein, while the second contains relatively pure $Cwp84_{33\text{-}700}$.

SAXS data were collected for constructs 1, 2, and 3, possessing zero, one, and two cell wall binding domains on beamline B21 at Diamond Light Source and preliminary analysis was performed with scatter (Robert Rambo, Diamond Light Source, unpublished). As would be expected, the data showed increasing D_{max} and R_g, confirming increasing particle size, however this was not further analysed, due to the publication of full length structures of Cwp6 and Cwp8 (Usenik *et al.*, 2017), which demonstrated flaws in constructs 2, 3, 5 and 6, in that they did not contain full cell wall binding domains. This is discussed further in Chapter 7.

4.3.15 Expression of constructs 5 to 7

Like constructs 2 to 4, constructs 5 to 7 showed a significant degree of degradation when expressed at 16 °C overnight which was reduced when the constructs were expressed for four hours at 37° C (figure 4.28). Unlike the large constructs, constructs 5 to 7 still expressed solubly at 37° C.

4.3.16 Purification of constructs 5 to 7

Attempts to purify the protein from constructs 5 to 7 were met with mixed success. The proteins were purified both with and without their GST tags to see if this reduced the degradation. Some degradation was observed for all constructs, with and without the GST tag, although the amount of degradation observed varied (figure 4.29).

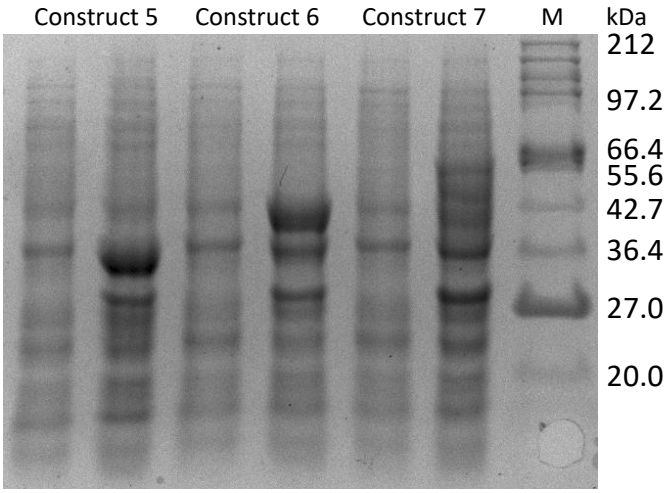


Figure 4.28 SDS-PAGE showing expression of constructs 5 to 7. Two samples are shown for each of the constructs, one taken just before induction, one taken just before harvesting. Construct 5 was predicted to be 36 kDa, construct 6 was predicted to be 47 kDa and construct 7 was predicted to be 58 kDa. Large bands are visible at the appropriate sizes for constructs 5 and 6, while construct 7 still showed a significant degree of degradation. More minor degradation can also be seen for constructs 5 and 6.

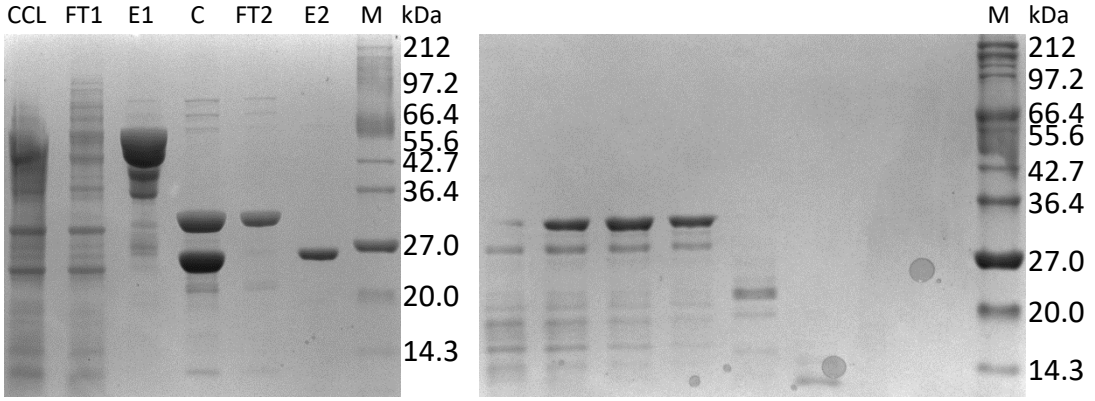


Figure 4.29 SDS-PAGE showing purification of construct 7. $Cwp84_{507\text{-}803}$ with GST tag – 58 kDa, $Cwp84_{507\text{-}803}$ without GST tag – 32 kDa. GST(left) and size exclusion (right). CCL – cleared cell lysate. FT1 – flow through from 1st round of GST purification. E1 – eluate from 1st round of GST purification. C – cleaved protein after dialysis with 3C protease. FT2 – flow through from 2nd round of GST purification with $Cwp84_{507\text{-}803}$. E2 – Eluate from 2nd round of GST purification. The largest band on the size exclusion gel corresponds to $Cwp84_{507\text{-}803}$, however some degradation can be seen. Other constructs showed similar or greater degrees of degradation.

4.3.17 Structural analysis of constructs 5 to 7

Despite some degradation, crystallisation screens were set up for constructs 5 to 7 both with and without GST tags. No crystals were observed. For constructs 5 and 6 this is likely to be due to the afore mentioned flaws in the constructs revealed by the recently reported full length structures of Cwp6 and Cwp8 (Usenik *et al.*, 2017).

4.3.18 Further evaluation of X-ray data

Much of the structural work presented here was performed some time ago, it was therefore decided that structural data should be reprocessed to ensure the best analysis of the data possible. Data were integrated with *DIALS* and scaled with *AIMLESS*.

Crystallographic statistics for the reprocessed data are summarised in table 4.30. Reprocessing resulted in significant improvements to the statistics for the structure with the propeptide, notably in the outer shell completeness. A significant increase in completeness was also observed for the first of the two structures without the propeptide, which also showed a significant degree of anisotropy that was not noticed when the structure was initially determined (figure 4.31). Very minor improvements were made to the second dataset without the propeptide upon reprocessing so this reprocessed dataset has not been included.

Table 4.30A Crystallographic statistics for Cwp84$_{33\text{-}497_C116A}$. Inner shell statistics are given in square brackets. Outer shell statistics are given in round brackets.

	Original data	Reprocessed data
Space group	$P2_1$	$P2_1$
Unit-cell parameters		
a (Å)	50.9	50.9
b (Å)	73.5	73.5
c (Å)	125.6	125.6
α = γ (°)	90.0	90.0
β (°)	93.6	93.6
Resolution range (Å)	[48.15 - 7.67] (1.42 - 1.40)	[50.81 – 7.67] (1.42 – 1.40)
R_{merge}	[0.138] 0.099 (0.258)	[0.122] 0.103 (0.900)
R_{meas}	[0.171] 0.121 (0.349)	[0.147] 0.125 (1.173)
R_{pim}	[0.058] 0.054 (0.301)	[0.081] 0.070 (0.568)
$CC_{1/2}$	[0.936] 0.989 (0.679)	[0.961] 0.994 (0.518)
I/σ(I)	[26.5] 16.0 (4.2)	[18.9] 10.3 (2.7)
Completeness (%)	[97.9] 93.9 (65.2)	[99.8] 99.1 (88.0)
Number of reflections	[5,982] 810,986 (14,198)	[7,162] 1,039,920 (30,743)
Unique reflections	[1,146] 170,213 (5,848)	[1,179] 180,011 (7,900)
Multiplicity	[5.2] 4.8 (2.4)	[6.1] 5.8 (3.9)
R_{work}/R_{free}	0.138/0.169	0.129/0.159
RMSDs		
Bond Lengths (Å)	0.008	0.013
Bond Angles (°)	1.340	1.644
Ramachandran Statistics (%)		
Favoured	96.1	96.8
Allowed	3.9	3.2
Outliers	0	0
Average B-factors (Å^2)		
Protein	16.7	15.3
Ligand	36.6	36.2
Water	29.8	25.7
Number of Atoms		
Protein	7404	7243
Ligand	104	63
Water	928	726
PDB code	4CI7	N/A

Table 4.30B Crystallographic statistics for Cwp84$_{92\text{-}497_C116A}$. Inner shell statistics are given in square brackets. Outer shell statistics are given in round brackets. The significant reduction in resolution resulted from the use of a more conservative cut-off due to anisotropy.

	Original data	Reprocessed data
Space group	P1	P1
Unit-cell parameters		
a (Å)	42.2	42.3
b (Å)	58.4	58.4
c (Å)	93.1	93.1
α (°)	89.3	89.3
β (°)	78.0	78.0
γ (°)	71.6	71.6
Resolution range (Å)	[55.35 – 9.01] (1.88 - 1.84)	[55.35 – 8.66] (2.61 – 2.50)
R_{merge}	[0.024] 0.071 (0.589)	[0.027] 0.053 (0.152)
R_{meas}	[0.033] 0.101 (0.833)	[0.038] 0.075 (0.215)
R_{pim}	[0.024] 0.071 (0.589)	[0.027] 0.053 (0.152)
$CC_{1/2}$	[0.994] 0.996 (0.610)	[0.999] 0.997 (0.981)
I/σ(I)	[40.3] 13.6 (2.7)	[30.2] 14.2 (5.4)
Completeness (%)	[82.3] 83.5 (83.0)	[99.3] 91.1 (93.3)
Number of reflections	[1,525] 155,876 (9,319)	[2,678] 99,899 (12,577)
Unique reflections	[508] 59,711 (3,659)	[679] 25,991 (3,265)
Multiplicity	[3.0] 2.6 (2.5)	[3.9] 3.8 (3.9)
R_{work}/R_{free}	0.227/0.291	0.185/0.249
RMSDs		
Bond Lengths (Å)	0.009	0.011
Bond Angles (°)	1.323	1.458
Ramachandran Statistics (%)		
Favoured	96.5	94.9
Allowed	3.5	5.0
Outliers	0	0.1
Average B-factors (Å^2)		
Protein	31.9	34.1
Ligand	31.6	33.5
Water	32.0	24.4
Number of Atoms		
Protein	6394	6341
Ligand	18	2
Water	449	85
PDB code	4D59	N/A

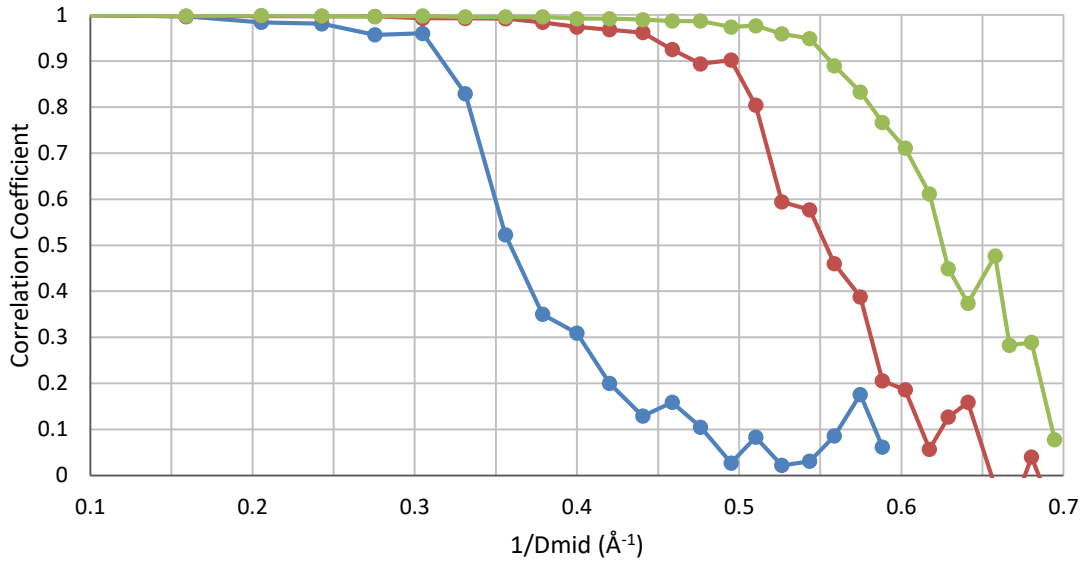


Figure 4.31 Anisotropy in the first structure of Cwp84$_{92\text{-}497_C116A}$. The correlation coefficient is given for each of the three axes in the dataset. If a $CC_{1/2}$ of 0.3 is taken as the point at which the resolution should be cut, it can be seen that in the worst direction the data extend to a 1/Dmid of 0.4 Å^{-1}, equivalent to a resolution of 2.5 Å, while in the best direction, they extend to a 1/Dmid of 0.66 Å^{-1}, equivalent to a resolution of 1.5 Å.

4.4 Discussion

4.4.1 Cysteine protease domain

The cysteine protease domain of Cwp84 assumes a cathepsin L-like fold with some notable differences. The protein possesses a slightly extended occluding loop region, lacks a prosegment binding loop (PBL) and has a somewhat deeper active site groove than normal. The catalytic residues Cys116 and His262 are found in a typical position at the top of the active site groove (figure 4.32). As predicted (Savariau-Lacomme *et al.*, 2003), Gln110 is nearby, able to assist in formation of an oxyanion hole, stabilising the catalytic intermediate. Asn294, however is not located within the active site so does not assist in catalysis.

The occluding loop is a feature found in cathepsin B-like cysteine proteases. The loop extends around the protein and partially blocks the active site. This confers a greater degree of substrate selectivity on the protein than is seen in cathepsin L-like cysteine proteases, while a conserved HH motif within the occluding loop confers

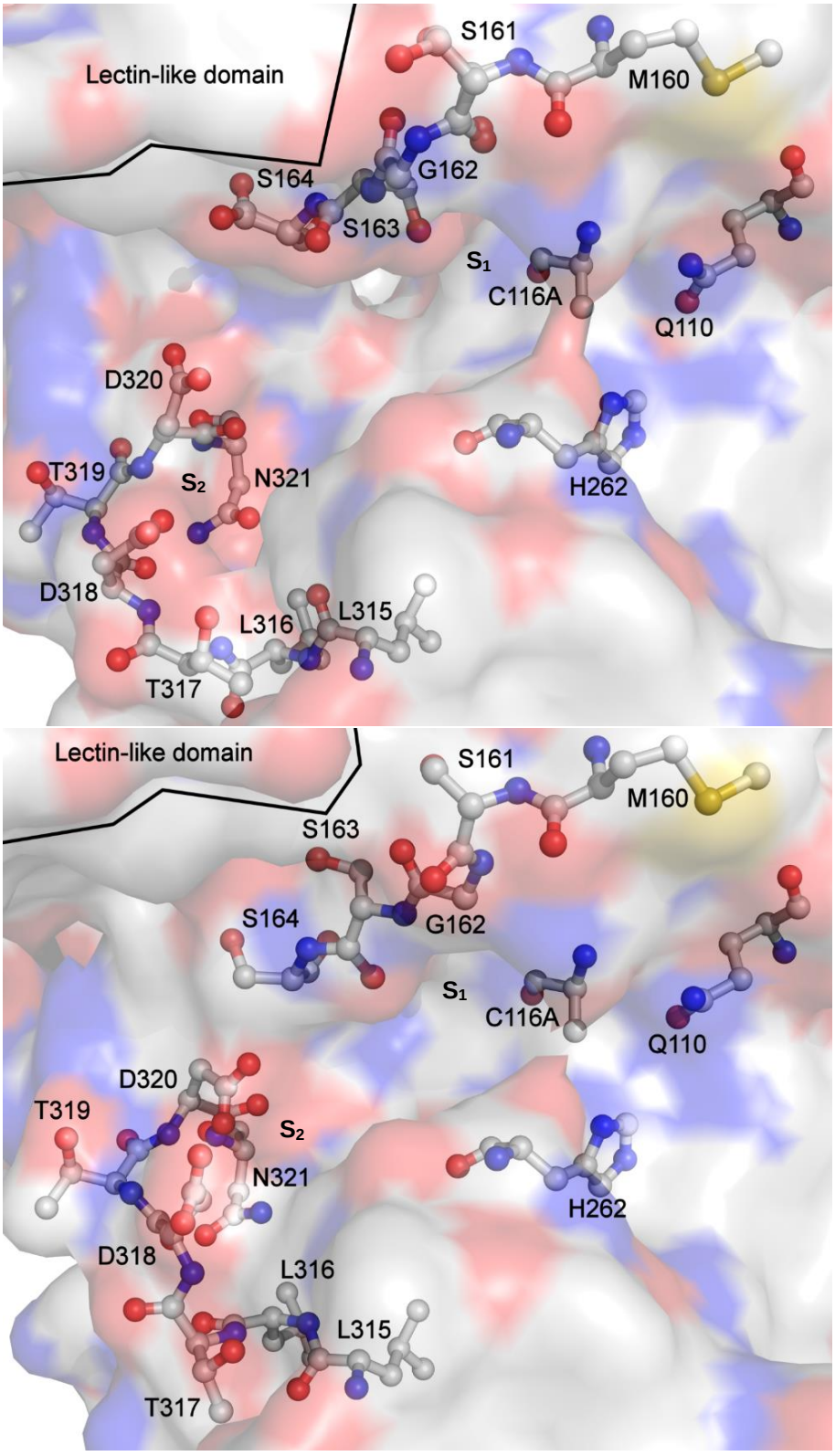


Figure 4.32 The active site of Cwp84 with (top) and without (bottom) the propeptide. Two conformational changes are noticeable upon propeptide cleavage: the rotation of the loop formed by Met160-Ser164, and the formation of a large negatively charged patch by Asp318 and Asp320. This is likely to be involved in substrate selectivity, binding the P_2 lysine of SlpA. The S_1 and S_2 binding pockets are indicated.

carboyxpeptidase activity. The equivalent loop in cathepsin L-like cysteine proteases is considerably shorter (Figure 4.33), so has very little effect on substrate binding or catalysis but has a well conserved fold (Sajid & McKerrow, 2002). In Cwp84, this loop is closer to that seen in cathepsin L-like cysteine proteases and does not contain the HH motif. Cwp84 does not, therefore, possess any carboxypeptidase activity. The loop is however, somewhat longer than is normally seen in cathepsin L-like cysteine proteases, so it is possible that it may still play a role in substrate binding (figure 4.33).

4.4.2 Propeptide

Cysteine protease propeptides in general have two widely observed important points of interaction with the mature cysteine protease domain: the prosegment binding loop (PBL) and the S_2 subsite (Sajid & McKerrow, 2002).

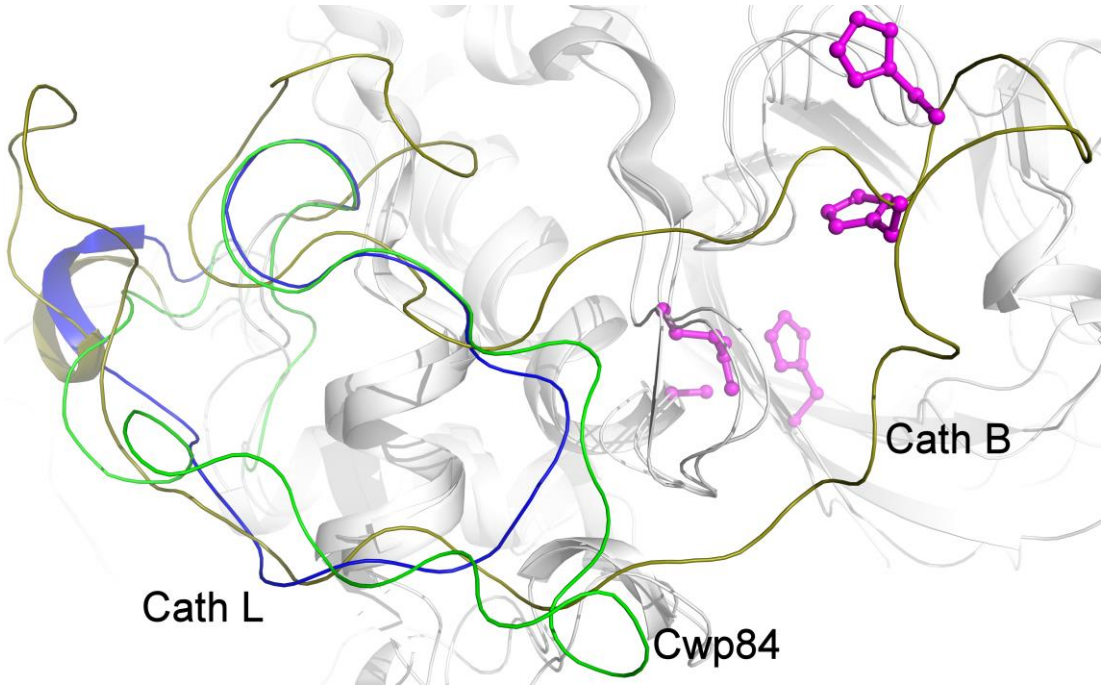


Figure 4.33 Cysteine protease occluding loops. The occluding loops of cathepsin L, in blue (1CJL), cathepsin B, in olive (1PBH), and $Cwp84_{33\text{-}497\_C116A}$, in green, are shown. The active site residues of Cwp84 and the HH motif of cathepsin B are shown in magenta. The occluding loop of Cwp84 is much closer to that of cathepsin L-like proteins, but is still significantly different to the usually well conserved conformation. It is possible that this loop may be involved in binding SlpA, but it is unlikely to play as much of a role in substrate binding as the occluding loop of cathepsin B-like proteins.

The propeptides of cysteine proteases usually wrap around the cysteine protease domain, stabilised by a peripheral portion known as the prosegment binding loop, forming a small β-sheet. As the propeptide in Cwp84 assumes a different conformation, wrapping around the adjacent lectin-like domain, Cwp84 does not possess a PBL and the fold of this part of the protein is significantly different to other cysteine proteases (figure 4.20).

As with other cysteine proteases, the propeptide of Cwp84 sits in the active site groove in the opposite direction to the substrate (figure 4.19). The S_2 subsite is a pocket in the active site groove to which the P_2 residue of the substrate, that is, the residue two before the scissile bond, binds, which the propeptide mimics. A residue positioned above this site has been shown to play an important role in substrate selectivity. In papain, this residue is a serine, in cathepsin L, it is an alanine, and in cathepsin B it is a glutamate. In Cwp84, with the propeptide bound, this region appears to form a negatively charged pocket containing Thr317, Asp318, Asp320 and Ser235. The P_2 residue of SlpA is usually a lysine, which will logically fit well into this pocket. Strangely, the residue located in approximately this position in the propeptide of Cwp84 is Val66 (figure 4.19). Despite this not mimicking the substrate, there is still a significant number of interactions between the propeptide and the cysteine protease in this region (figure 4.34). Indeed, upon propeptide cleavage, the S_2 pocket is one of three parts of the protein that show significant conformational changes.

In the presence of the propeptide, Thr319 hydrogen bonds to Met73 and Arg215. When the propeptide is removed, this stabilisation is lost and a loop formed by Leu315-Asn321 moves away from the positions of the other two residues, closer to the active site groove, including a 4 Å movement of Asp318 (figure 4.32). This forms a negatively charged patch with Asp318 and Asp320, potentially allowing better stabilisation of the P_2 lysine of SlpA. Alternatively, this may simply demonstrate flexibility in this loop, which may undergo another conformational change upon substrate binding.

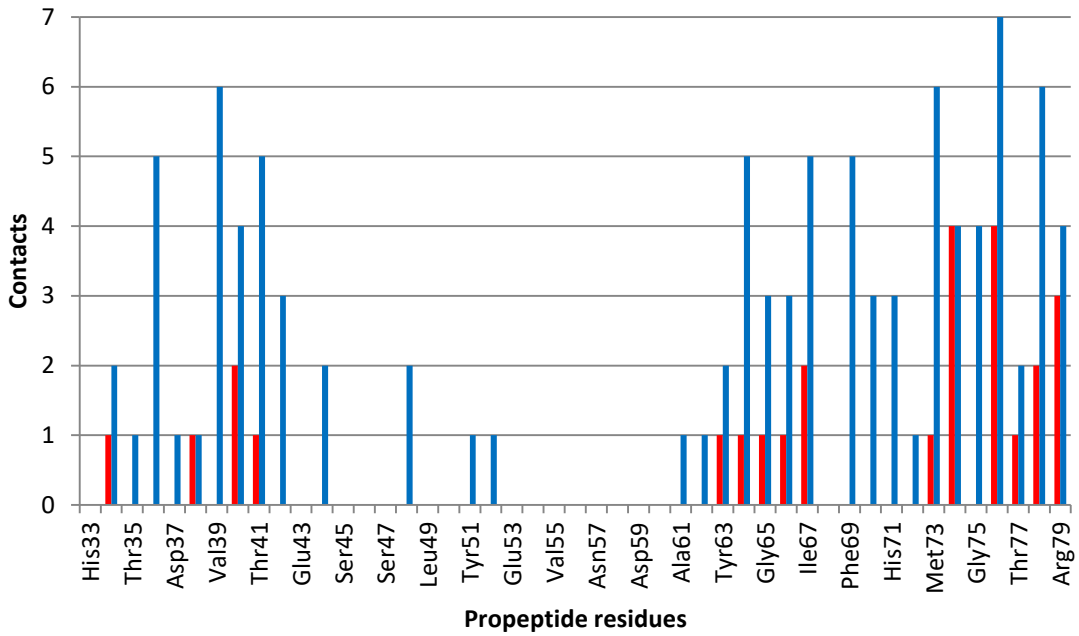


Figure 4.34 Contacts between the propeptide and the cysteine protease or lectin-like domains. Hydrogen bonds (Charge based interactions up to 3.2 Å) are indicated by red bars, Van der Waals contacts (any interactions up to 4.2 Å) are indicated by blue bars. The three peaks, which are particularly noticeable when examining the hydrogen bonds, closely correlate with the three loops that undergo conformational changes upon propeptide cleavage.

A second conformational change is observed in the S_1 pocket - Met160-Ser164 forms part of a hydrogen bond network to the propeptide, which also involves Tyr63-Val66 in the propeptide, Asn114 in the cysteine protease domain, and Tyr455 in the lectin-like domain. Most notably, Tyr455 loses hydrogen bonds to Tyr63 and Gly162. This appears to cause the backbone of Met160-Ser164 to rotate, including a rotation of approximately 160° in the peptide bond between Ser161 and Gly162, with the carbonyl now pointing away from the active site groove, rather than toward it. This is accompanied by a movement of the loop formed by these residues the active site residues, with a 3 Å movement of the alpha carbon of Gly162 (figure 4.32). This movement, which enlarges the P_1 pocket and reduces its polarity, may aid substrate binding of SlpA as the P_1 residue is usually alanine.

Comparing these two conformational changes to archetypal C1A cysteine proteases, the equivalent loops in papain (Kamphuis *et al.*, 1984; Roy *et al.*, 2012), cathepsin L (Coulombe *et al.*, 1996; Adams-Cioaba *et al.*, 2011), and cathepsin B (Musil *et al.*,

1991; Turk *et al.*, 1996; Podobnik *et al.*, 1997) do not appear to undergo any conformational changes upon propeptide cleavage. This suggests the significance of these changes, which are likely to facilitate the binding of SlpA to the active site.

4.4.3 Lectin-like domain

Upon initial visual inspection of the newly identified domain between the cysteine protease domain and the first cell wall binding domain, it was observed that the domain bears a similar fold to that seen in lectin domains. This was confirmed by a DALI search against the domain as many of the highest hits were lectins. Blast searches against Cwp84 also revealed a range of other cysteine proteases with apparent lectin-like domains, some of which possess cell wall binding domains while others do not (figure 4.21). Could the function of the lectin-like domain be closely coupled to the function of the cysteine protease domain? At the very least, it seems likely that the positioning of the lectin-like domain in very close proximity to the active site (figure 4.19) will have an effect on catalysis. Understanding the role of the lectin-like domain, therefore seems important to understanding how Cwp84 functions.

Lectins, derived from the Latin *lectus*, meaning “picked”, “chosen”, or “selected” (Boyd, 1954), are a large family of carbohydrate binding proteins. Initial work focused mainly on plant lectins, with the potent ribonuclease and biological weapon ricin being the first to be identified, although hundreds of different lectins have since been shown to be expressed by species from all kingdoms (Sharon & Lis, 2004). Carbohydrate specificity was first demonstrated for concanavalin A, which binds to glycans on the surface of red blood cells causing haemagglutination – the clumping of red blood cells. This effect is inhibited by sucrose (Sumner & Howell, 1936; Sharon & Lis, 2004). Different lectins are capable of binding to anything from simple sugars to much more complex glycoconjugates, although specificity for a particular carbohydrate is a commonly shared feature (Coelho *et al.*, 2017).

Lectins play roles in a wide range of biological processes including cell-cell interactions, recognition of secreted molecules, binding to cells by secreted molecules, host cell invasion, immunomodulation and proliferation (Sharon & Lis,

2004; Coelho *et al.*, 2017). Despite this, their presence in a very large number of different organisms, the diverse range of carbohydrates targeted and a lack of sequence similarity, lectins tend to share a remarkably well conserved fold. Specifically, this fold consists of two antiparallel β-sheets, which come together to form a β-sandwich or jelly roll (Sharon & Lis, 2004).

C-type lectins are a group of lectins found in animals, named for the calcium dependence of those first identified. They were initially classified into seven groups (Drickamer, 1995), but this has since been expanded upon, and currently consists of seventeen (Zelensky & Gready, 2005). Despite the name, many proteins with a C-type lectin fold have been shown not to bind calcium.

The fold of the lectin-like domain of Cwp84 is stabilised by the presence of a calcium ion near to the interface with the cysteine protease domain (figure 4.35). This calcium ion is coordinated by the carbonyls of Leu339 and Lys460 and the side chains of Glu448 and Asn487. Leu339 and Lys460 show little to no conservation while Glu448 and Asn487 appear moderately well conserved, despite a low degree of conservation in surrounding residues (figure 4.21). This suggests that related proteins may also bind calcium and, due to the proximity of the calcium binding site to the active site, that calcium binding may be important for catalysis. This conservation is not complete, however, particularly in the case of Asn487. Mutations of Glu448 and Asn487 could be used to determine the importance of calcium binding to both correct folding of Cwp84 and catalysis.

Ligand binding modes can vary significantly, as can specificity (Drickamer, 1999). This can make determination of the potential ligand of a lectin and elucidation of the function somewhat difficult. This is the case for Cwp84. Based on structural similarity, this domain is referred to as a “lectin-like” domain, and notably, it is stabilised by a calcium ion, however the ability to bind any carbohydrates has yet to be demonstrated.

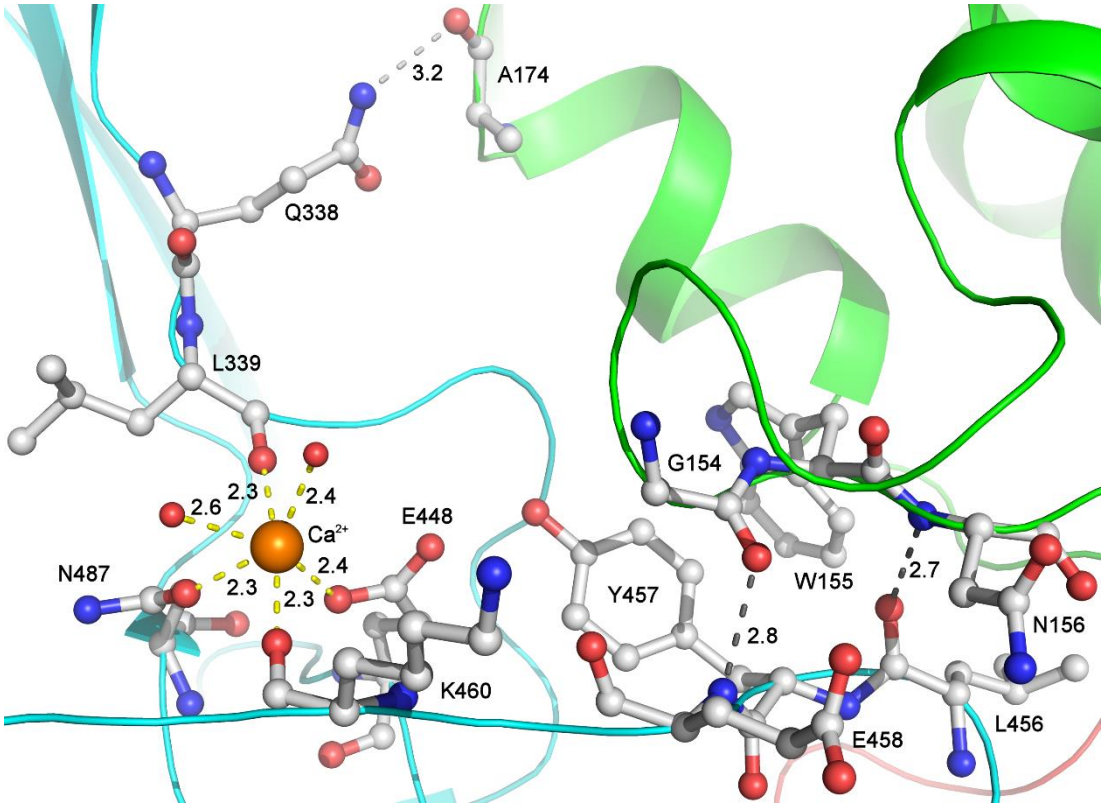


Figure 4.35 Location of the calcium ion in the lectin-like domain. The calcium ion is coordinated by four residues from different points in the primary sequence of the lectin-like domain. This is likely to stabilise the fold of the domain. This portion of the lectin like domain also interacts closely with the cysteine protease domain near to the active site – the propeptide can be seen in red in the active site to the bottom right of the image.

The lectin-like domain contains a hydrophobic core, which opens partially on the surface of the protein, producing a hydrophobic pocket formed by Ile347, Ile468, Ile477 and Phe483. Leu36 and Val39 from the propeptide insert into this pocket, potentially stabilising the fold through hydrophobic interactions. Lys34 stabilises the conformation through a hydrogen bond to Thr479.

This hydrophobic pocket is subject to the third observed conformational change that occurs upon propeptide cleavage. Thr479-Pro485 form a loop on the top of the hydrophobic pocket. This loop assumes two different conformations in the structures presented here. In the first structure without the propeptide, aside from a slight movement away from the position of the propeptide, the conformation is largely

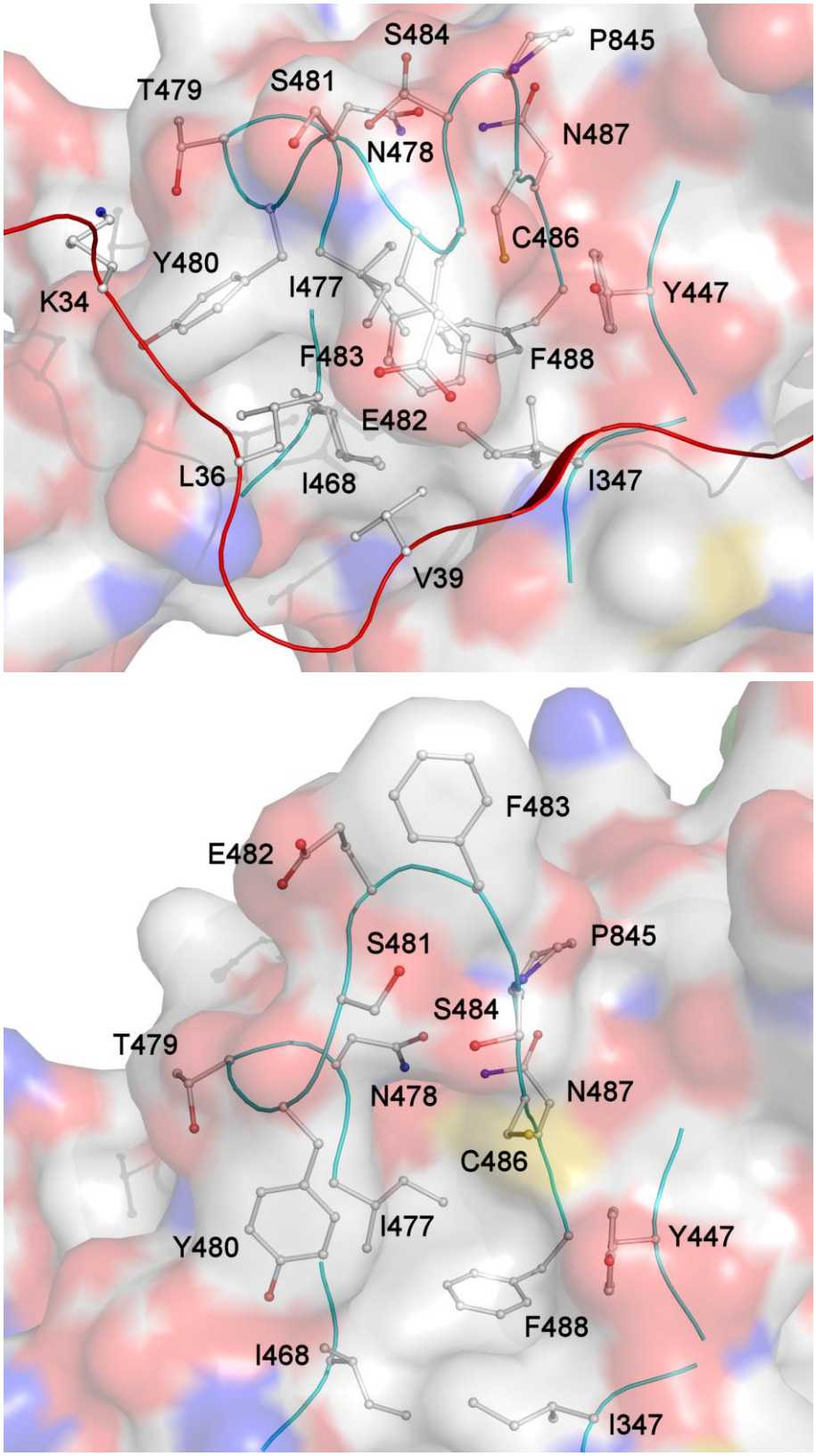


Figure 4.36 The lectin-like domain hydrophobic pocket. The propeptide (top) is anchored to the lectin like domain by Lys34, Leu36 and Val39. Upon cleavage of the propeptide the Thr479-Pro485 loop is no longer stabilised by this anchoring, so is able to swing open, exposing the hydrophobic pocket on the surface of the protein.

unchanged, while in the second structure, the loop assumes a markedly different conformation, making the hydrophobic pocket much more accessible (figure 4.36).

This conformation does appear to be partially stabilised by symmetry contacts but it seems that it is only possible for this conformation to be present because of the loss of the stabilisation from the propeptide. At the very least, the cleavage of the propeptide results in increased flexibility in this loop and a light increase in solvent exposure of the hydrophobic pocket.

Three significant conformational changes have been noted that occur upon propeptide removal. Met160-Ser164 and Leu315-Asn321, which form part of the active site groove, and Thr479-Pro485 on the surface of the lectin like domain. Figure 4.34 shows a count of interactions between the propeptide and residues of either the cysteine protease domain or lectin-like domain. Three peaks can be clearly seen on this graph that closely correlate with the portions of the propeptide that interact with the regions that undergo conformational changes, confirming that these conformational changes do occur as a result of the loss of the propeptide.

4.4.4 Inhibitor and substrate binding

Attempts were made to co-crystallise $Cwp84_{92\text{-}497\_C116A}$ with the cysteine protease inhibitor E-64 and peptides based upon the cleavage site from SlpA. Although crystals that diffracted well were obtained, nothing was observed to be bound to the active site. E-64 has previously been demonstrated to be capable of inhibiting Cwp84 moderately well (Janoir *et al.*, 2007). However, as previously discussed, the binding of E-64 to cysteine proteases is primarily dependant on interactions between the cysteine protease and the epoxysuccinyl group, with binding being made irreversible by the formation of a covalent bond to the active site cysteine (Matsumoto *et al.*, 1999). As the active site cysteine is mutated to an alanine in the present construct, it is understandable that binding may not have been strong enough for the formation of a complex.

In the case of the substrate peptides, the failure to form a complex is likely to be to do with how SlpA binds to Cwp84. The two peptides used were LETKS|ANDTI (strain

630) and RUTTKS|AAKASI (ribotype 027). Apart from the KS|A cleavage site, the two peptides possess very little similarity, and even these three residues are not conserved in some other strains. Sequence comparisons of SlpA from a range of strains have demonstrated a lower degree of similarity than might be expected (Calabi & Fairweather, 2002). Despite the differences in cleavage sites, only two residues on the surface of the lectin-like domain are different in Cwp84 between strain 630 and ribotype 027. This may demonstrate that the binding of SlpA to Cwp84 is considerably more complex than a simple interaction at the active site. It has previously been discussed that the occluding loop of cathepsin B-like cysteine proteases has a role in substrate binding, while the well conserved equivalent loop in cathepsin L-like cysteine proteases does not. This work has shown that the occluding loop in Cwp84 assumes a somewhat different fold. Could this be required to facilitate SlpA binding? Notably, figure 4.21 demonstrates that this alternative occluding loop shows little to no conservation. Could this be an indicator that the occluding loop changes with the substrate, or does it indicate that the occluding loop is unimportant? Clearly, the structure of a complex between Cwp84 and SlpA is required to answer these questions. Formation of a complex may also be required to stabilise the portion of SlpA that is yet to have its structure determined (Fagan *et al.*, 2009) (potentially equivalent to domain 3 in Cwp2 and Cwp8, discussed in Chapter 6).

4.4.5 Lectin-like domain ELISA

An ELISA was performed to determine potential ligands for the lectin-like domain. Rather than having an intention of determining the actual ligand that the lectin-like domain binds normally, the purpose was to confirm the ability of the lectin-like domain to bind carbohydrates, any positive results would have been followed up with ITC and crystallisation experiments to further characterise binding. Unfortunately, the assay did not work. No difference was seen for each of the sugars, so if they were able to bind to the plate, they did not affect lectin binding. The increased signal strength of Concanavalin A over Cwp84 both with and without the propeptide and of Cwp84 over the blank, suggest that the proteins did bind to the plate, either via the

bound carbohydrates or directly to the plate, which would suggest that blocking did not work.

As mentioned in section 4.4.3, lectins are capable of binding a wide range of carbohydrates, even if this assay had worked, there is no guarantee that $Cwp84_{33\text{-}497\_C116A}$ or $Cwp84_{92\text{-}497\_C116A}$ would have bound any of the carbohydrates present. Identification of a ligand is likely to require screening via glycan arrays involving hundreds of different carbohydrates. The possibility of using such arrays has been explored, but they are yet to be performed.

4.4.6 Further evaluation of X-ray data

As the data and resulting structures presented in this Chapter were originally processed and published some time ago, it was decided that they should be reprocessed to ensure that the best possible analysis could be performed and so that the data could be presented in the best way possible. One notable, somewhat worrying statistic that drove this was the completeness of the lower resolution structure without the propeptide but it was decided that all three structures should be examined.

The main effect that reprocessing of the data for the structure with the propeptide had is the detection of a significant amount of weak spots that were not originally included. The best way to handle weak data has been the subject of a significant amount of discussion in recent years and has led to improvements to software in both the detection and the use of weak data (Evans & Murshudov, 2013; Waterman *et al.*, 2016). This resulted in the total number of observed reflections increasing by over 28%, the total number of unique reflections increasing by nearly 6% and significant improvements to the completeness and multiplicity in all resolution shells. That the extra data consists of weak reflections is demonstrated by increases in R_{merge}, R_{meas} and R_{pim} in the majority of shells and decreases in $I/\sigma(I)$. Increases in the inner shell and overall $CC_{1/2}$, indicate, however that the data have been improved by the inclusion of weak reflections. The outer shell $CC_{1/2}$ decreased with the inclusion of weak data, however this is offset by a significant increase in completeness. Improvements to the R_{work} and R_{free} and to the Ramachandran statistics may indicate

that the structure better reflects the data due to the increase in completeness and multiplicity, particularly at high resolution, however this may also be a result of slightly looser restraints during refinement in line with recent publications on the matter (Jaskolski *et al.*, 2007; Moriarty *et al.*, 2016; Wlodawer, 2017), reflected by slightly higher RMSDs.

An alignment was performed between the reprocessed structure and the published structure to determine how much they differ from each other. A very low RMSD of 0.04 Å (5768 atoms) was observed between the two versions of the structure with the propeptide, indicating that the structures are largely identical, this was confirmed by a residue by residue comparison of the two structures, which resulted in very few noticeable differences. This demonstrates that although the structure may be arguably statistically better, the resulting biological information is completely unchanged. Because any differences between the structures are extremely minor, the published structures and the resulting conclusions are still completely valid (Matthew Conroy, PDBe, personal communication).

The significant degree of anisotropy observed in the data for the first structure without the propeptide lead to a different strategy of cutting the data: conservatively, based upon usable diffraction in the weakest direction. While this makes direct comparison of the statistics more difficult, it can be seen that the completeness and multiplicity have been improved at all resolutions. This is likely to be due to the inclusion of weaker data, which is potentially evidenced by the significant decrease in the inner shell I/σ(I), despite only a slight increase in resolution in this shell. The somewhat poorer Ramachandran statistics may reflect the loss of good data, but the significantly improved R_{work} and R_{free} are likely to reflect the removal of unusably noisy data.

The original structure and the reprocessed structure without the propeptide superpose with an RMSD of 0.18 Å (5,377 atoms), this is considerably larger than for the structure with the propeptide, but still easily acceptable and considerably less than the RMSD that might result from the same structure determined from different crystal forms - up to approximately 0.5 Å (Chothia & Lesk, 1986). The increase relative

to the structure with the propeptide may result from the slightly different cell dimensions and the significant increase in completeness. A visual inspection between the two structures resulted in very few differences being observed.

On the wider subject of anisotropy, there is currently a significant degree of debate within the crystallographic community about how it should be handled. Throughout the history of crystallography, the maximum resolution of a dataset has been cut spherically. If the data extend to approximately an equal resolution in all directions, this is a perfectly valid way of handling the data, however, in reality, all data exhibit a degree of anisotropy, except those in cubic space groups (Evans & Murshudov, 2013). If the level of anisotropy is particularly high, cutting the data spherically may result in the inclusion of a large amount of noisy data on at least one axis which can result in a noisier map. A model may be built into the errors in this noisier map, resulting in a less accurate structure with a correspondingly higher R_{free}. Alternatively, the data may be cut conservatively to avoid the bad data, but this results the loss of good data on at least one axis. This may result in features of the map being missed. In reality, a spherically cut dataset is likely to include some bad data and miss some good data. Strategies have been developed to assess the degree of anisotropy (Evans & Murshudov, 2013) and to partially correct for it (McCoy, 2007), but they do not eradicate the problem. A strategy has recently been developed by Global Phasing (Bricogne *et al.*, unpublished, staraniso.globalphasing.org), that cuts the data ellipsoidally using an I/σI cut-off. This method is, however, yet to gain widespread usage and ellipsoidally cut data does not yet seem to be handled properly by refinement programs. Wider use and refinement of this method should result in slight improvements to datasets and resulting structures. Although this is very unlikely to result in any significant changes to biological conclusions, it may well be the difference between being able and being unable to solve a particularly difficult structure.

4.4.5 Conclusions

This work has resulted in the determination of the structure of the cysteine protease domain and the newly discovered lectin-like domain of Cwp84 both with and without

the propeptide. The propeptide sits in the active site groove and, in a novel fold, wraps around the lectin-like domain, a feature that is likely to be important to the initial folding of both domains. The cysteine protease domain is similar to other cathepsin L-like cysteine proteases but bears some notable structural differences, specifically, the lack of a prosegment binding loop, an alternative occluding loop structure and a deeper active site groove, due to the presence of the lectin like domain.

Upon propeptide cleavage, three conformational changes have been observed, two of which are likely to be required to assist in binding of SlpA to the active site, while the third is on the surface of the lectin-like domain.

The role of the lectin-like domain and any ability to bind carbohydrates was explored through a carbohydrate ELISA but this was unsuccessful. Further work is therefore required to determine the function of the lectin-like domain and ligands it is able to bind. Proteins containing a cysteine protease domain and a lectin like domain appear to be produced by a wide range of organisms, so it is likely that the lectin-like domain possesses a conserved unknown function in all of these proteins.

Work was also performed on constructs containing cell wall binding domains, which suffered from degradation issues. As the constructs did not code for complete cell wall binding domains, it is likely the proteins were misfolded. Full length structures of Cwp6 and Cwp8 have since been published (Usenik *et al.*, 2017).

The structures of Cwp84 presented here go some way to improving understanding of the role of Cwp84 in the cleavage of SlpA and the formation of the S-layer of *C. difficile*. Some work has been done to design specific Cwp84 inhibitors (Dang *et al.*, 2010; Gooyit & Janda, 2016), but this area is yet to be fully explored and the structures presented here will be vital to this process in the future if Cwp84 is to be used as a drug target to fight *C. difficile* infections.

Chapter 5

Cwp19

5.1 Introduction

The gene coding for Cwp19 is found within the AP locus (figure 1.3). The AP locus contains several genes required for the correct formation and secretion of PSII, the repeating hexasaccharide that mediates binding of the three cell wall binding domains of the Cwps to the cell wall (Ganeshapillai *et al.*, 2008; Willing *et al.*, 2015; Chu *et al.*, 2016). It is therefore possible that Cwp19 is involved in this process. The majority of genes within the AP locus are predicted to code for glycoside hydrolases (Sebaihia *et al.*, 2006; Monot *et al.*, 2011).

Cwp19 has been determined by Pfam and BLAST to contain an N-terminal glycoside hydrolase-like 10 (GHL10) domain with a high degree of certainty ($E = 10^{93}$). Pfam also gives a potential classification within the same region of a family 27 glycoside hydrolase (GH27), while a BLAST search also suggests a GH36. Both of these classifications, however, have much lower degrees of certainty than GHL10 (Altschul *et al.*, 1990; Eddy, 2008; Finn *et al.*, 2016).

Glycoside hydrolases (GHs) are present in virtually all organisms, with the only exceptions being a small number of archaea and some single celled eukaryotes. They are defined as catabolic enzymes responsible for the cleavage of O-glycosidic bonds, ie. the breakdown of carbohydrates (Naumoff, 2011b), well known examples of GHs are the oft-studied lysozyme and amylase. The carbohydrate active enzyme database (CAZy) originally classified GHs into 35 families based on their primary sequences, they are now divided into 14 clans, which are subdivided into approximately 140 families, however this number is ever increasing (Henrissat, 1991; Lombard *et al.*, 2014).

Membership of a clan is determined based on a conserved structure, mechanism and substrate specificity. The determination of a substrate for a specific glycoside hydrolase, whether the enzyme follows a mechanism that retains or inverts anomers and ultimately the function of the GH of interest, is frequently problematic as many glycoside hydrolases share common folds, with substrate specificity and mechanism determined by subtle differences within the active site (Naumoff, 2011b). It is believed that all members of a given clan are divergently evolved.

GHL1-GHL15 were identified in 2011 as families of proteins that are likely to possess at least some glycoside hydrolase activity and have the common $(\beta/\alpha)_8$ motif, which forms an eight stranded β-barrel known as a TIM barrel (named after triose phosphate isomerase, the first structure to be determined with said fold (Alber *et al.*, 1981)) (Naumoff, 2011a). GHL1 has since been reclassified as GH129 (Kiyohara *et al.*, 2012), while the remaining GHL families are yet to be characterised. GHs with TIM barrel folds fall into clans GH-A, GH-D, GH-H and GH-K.

Pfam reports that over 1000 protein sequences have been classified as containing GHL10 domains, of these, around 1% have been identified in fungi and animals, while the remaining 99% are spread across a wide range of bacterial phyla (Finn *et al.*, 2016).

Polysaccharides can essentially be divided into two groups based on their functions: energy storage and cellular structure. Glycoconjugates, on the other hand, usually have higher order functions such as cell to cell interactions and modulation of activity. The large amount of isomerism exhibited by monosaccharides, coupled with a broad range of potential linkages, necessitates a great deal of diversity among enzymes that process carbohydrates, including glycoside hydrolases (Fushinobu *et al.*, 2013).

It has been noted that the 30 GHs identified in *Mycobacterium tuberculosis* can be categorised into four broad functional groups: metabolism of α-glucans produced by the bacterium, peptidoglycan maintenance, hydrolysis of β-glucans (primarily those consumed by the host), and α-demannosylation of proteins produced by the bacterium as a method of functional modulation (van Wyk *et al.*, 2017). It stands to reason that the majority of GHs in other bacteria may also fit into these four categories. If Cwp19 does have a role in processing surface exposed polysaccharides, as the localisation of the gene suggests, this would place it into the first two of the four potential functional categories.

The gene coding for Cwp19 has been shown to be present with more than 95% amino acid sequence identity in a wide variety of *C. difficile* strains (Biazzo *et al.*, 2013; Chu *et al.*, 2016). Although the expression of Cwp19 has yet to be thoroughly analysed, it

is known to be present in the S-layer under at least some conditions as it co-purified with Cwp84 in a pull-down assay using probes based on E-64, a cysteine protease inhibitor (Dang *et al.*, 2010).

A recent study on seven *C. difficile* strains found in Brazil indicated that the amount of Cwp19 in S-layer extracts was higher than any other protein in three strains and second only to Cwp2 in two strains and SlpA in one (Ferreira *et al.*, 2017). Such an apparently high level of expression would suggest a very important role for Cwp19, demonstrating the importance of characterisation of the protein.

A construct coding for Cwp19 from *C. difficile* strain QCD32g-58 minus the signal peptide and cell wall binding domains (residues 27-401) was previously cloned into the pET28a kanamycin resistant hexahistidine tagged vector by Jon Kirby (Kirby *et al.*, 2011). Expression, purification and crystallisation protocols were determined, which resulted in crystals that diffracted to approximately 2.0 Å, however molecular replacement was unsuccessful. It was suggested that this may be due to a range of defects with the data collected (Kirby *et al.*, 2011). It was therefore decided that the protocols should be re-established with the hope of producing crystals suitable for the determination of the structure of Cwp19.

5.2 Methods

5.2.1 Expression and purification

The same construct as previously used by Jon Kirby, coding for $Cwp19_{27\text{-}401}$, was transformed into *E. coli* BL21 Codon-plus cells and expressed overnight as described in the general methods. Cell pellets were resuspended in lysis buffer (25 mM Tris, 200 mM NaCl, 40 mM imidazole, pH 8.0), lysed at 20 KPSI in a French press. Lysate was cleared by centrifugation and $Cwp19_{27\text{-}401}$ was purified using a nickel affinity chromatography column, eluting with a single step increase in imidazole concentration to 200 mM. The imidazole was removed using a desalting column. Attempts were made to confirm the identity of the purified protein by mass

spectrometry but transferring the protein to pure water or 0.1% acetic acid invariably resulted in degradation, so mass spec was not performed.

Selenomethionyl-protein was produced by inhibiting methionine production as described in the general methods. Buffers used for IMAC had 2 mM DTT added, while the desalting buffer had 5 mM reduced glutathione added to prevent loss of anomalous signal through oxidation (Walden, 2010).

5.2.2 Crystallographic studies

To avoid previously identified issues with data that may have resulted in the unsuccessful molecular replacement (Kirby *et al.*, 2011), crystallisation conditions were re-screened at a range of protein concentrations as described in the general methods. Crystals grown in Molecular Dimensions Heavy and Light (H&L) condition H11 (50 mM KH_2PO_4, 14 % PEG 8,000) diffracted to 2 Å with a degree of anisotropy. The identified condition was screened around and supplemented with a range of other screens at a concentration of 10%. The Final crystallisation conditions are given in section 5.3.2.

Crystals were cryo-protected by addition of PEG 8,000 to a final concentration of 35-40%. High resolution native data were collected on beamline I02 at Diamond Light Source, while Se-MAD data were collected on I04 using the mini-kappa goniometer to maximise anomalous signal (Flaig *et al.*, 2013). Several datasets were collected at the peak and inflection energies of the Se-K edge and at high and low remote energies for structure determination by MAD.

Data were indexed and integrated with *XDS* (Kabsch, 2010) using the Xia2 pipeline 3dii (Winter, 2010; Winter *et al.*, 2013). The three integrated datasets were scaled together with *XSCALE* (Kabsch, 2010), before merging with *AIMLESS* (Evans & Murshudov, 2013). A high resolution cut-off was selected based upon an anomalous correlation coefficient of 0.3. The merged data were fed into the CRANK2 pipeline (CCP4, 1994; Skubak & Pannu, 2013) using SFtools, SHELXC and D (Sheldrick, 2008), REFMAC5 (Murshudov *et al.*, 2011), MAPRO, Solomon (Abrahams & Leslie, 1996), Multicomb, Parrot (Cowtan, 2010) and *Buccaneer* (Cowtan, 2006). Model building

was completed and the structure was refined with *Coot* (Emsley & Cowtan, 2004) and REFMAC5.

The high resolution data were indexed and integrated with *DIALS* (Waterman *et al.*, 2016), the number of observed reflections in the dataset mandated that this be done on a computer cluster, particularly dials.refine, which required more than 128 GB RAM. The data were scaled with *AIMLESS*, with a high resolution cut-off determined based on an anisotropic $CC_{1/2}$ of 0.3. Refinement was attempted at higher resolution, but this resulted in significantly higher R-factors and a noisier map. The Se-SAD structure was used as a model for molecular replacement with Phaser (McCoy *et al.*, 2007), the output of which was again refined using *Coot* and REFMAC5. Geometric restraints were relaxed somewhat relative to those recommended by Engh and Huber (Engh & Huber, 1991) based on recommendations by Jaskolski et al. (Jaskolski *et al.*, 2007). Phenix (Adams *et al.*, 2010) was used to refine occupancies. The structures were validated with MolProbity (Chen *et al.*, 2010).

5.2.3 Peptidoglycan hydrolase assays

20 mg of Lyophilised *Micrococcus luteus* cells (Sigma Aldrich) and a complete EDTA-free protease inhibitor tablet (Roche) were resuspended in 40 ml of 40 mM citrate, 40 mM K_2HPO_4 with the pH adjusted to a range of values between 3.9 and 6.6 with KOH. Approximately 600 µl of cellular suspension was diluted by addition of approximately 1.9 ml of buffer to a volume of 2.5 ml and a target OD_{450} of 0.6 to 0.65 (measured at 0.621 ± 0.032 (SD) across all samples). Samples in a quartz cuvette were heated to 37 °C and stirred in a nanodrop 2000c (ThermoFisher) and covered with parafilm to reduce evaporation. The parafilm was pierced and 100 µl of $Cwp19_{27\text{-}401}$ at 5.0 mg ml^{-1} was added. The OD_{450} was measured approximately every 2 seconds over the space of 2 hours. Each pH was performed three times and a control without addition of $Cwp19_{27\text{-}401}$ was performed at each pH. The rate of reaction was assessed by calculating the change in OD_{450} over the first three minutes and over the two hours as a proportion of the starting OD_{450} minus the change in OD_{450} of the control. Each calculation used the average of five time point measurements to reduce noise. A positive control was also performed with lysozyme at pH 6.2 (Shugar, 1952).

5.2.4 Benedict's assay

Solutions of 11 carbohydrates ranging from disaccharides to polysaccharides were produced at final concentrations of 0.25% for reducing sugars and agarose and 0.5% for non-reducing sugars. 500 µL of each solution was incubated for four hours at 37 °C with $Cwp19_{27\text{-}401}$ at 200 µg ml^{-1} and without $Cwp19_{27\text{-}401}$. After incubation, 500 µL of Benedict's reagent was added and samples were incubated at 95 °C for 10 minutes. The absorbance of each sample was measured at 320 nm to determine the extent of copper reduction. Six replicates were measured for each carbohydrate with and without $Cwp19_{27\text{-}401}$. A decrease in A_{320} relative to the control samples was taken as an increase in reducing ability of the sample, which was interpreted as an indicator of the ability of $Cwp19_{27\text{-}401}$ to hydrolyse at least one type of glycosidic bond in the sample. The breakdown of starch by amylase was used as a positive control (n = 3). Samples (n = 3) of glucose and maltose at a range of concentrations were used to determine an appropriate concentration for the reaction and to demonstrate that it is possible to use Benedict's test to differentiate between a monosaccharide and a disaccharide.

5.2.5 Substrate docking

SwissDock (Grosdidier *et al.*, 2011) was used to model N-acetylglucosamine, maltose, lactose, cellobiose and melibiose into the high resolution structure of $Cwp19_{27\text{-}401}$. The program was run with the most thorough settings, allowing flexibility in side chains up to 5 Å from the ligand.

5.3 Results

5.3.1 Expression and Purification

Cwp19$_{27\text{-}401}$ could be purified to a high degree using a single step IMAC protocol. Running the protein on a highly overloaded polyacrylamide gel revealed only one minor contaminant (figure 5.1).

5.3.2 Crystallisation and Structure Determination

Crystals were observed in two conditions – heavy and light (H&L) B11 (50 mM monobasic potassium phosphate, 20% PEG 8,000) and H11 (50 mM monobasic potassium phosphate, 14% PEG 8,000). A single large crystal grew in the latter condition with a lower PEG concentration, while much smaller, more fragile crystals grew in the former (figure 5.2).

Data were collected from the single large crystal, however, when processed they showed a moderately high anisotropic delta B of 19.6 Å^2. Screens around these

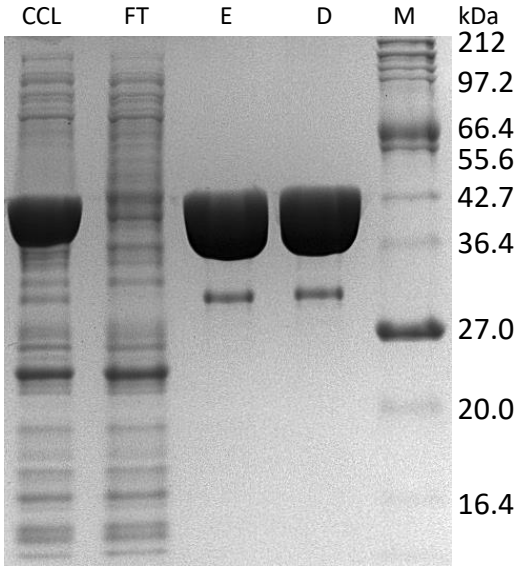


Figure 5.1. SDS-PAGE showing purification of Cwp19$_{27\text{-}401}$. The purified protein was predicted to have a mass of 45 kDa. CCL – Cleared cell lysate. FT – flow through from nickel column. E – Eluate from nickel column. D – Desalted protein. A good level of purity was achieved for Cwp19$_{27\text{-}401}$ with a single step. The desalting step was included to remove imidazole from the sample. The minor contaminant visible in the eluate and desalted samples could not be removed but was present in levels that did not prevent crystallisation.

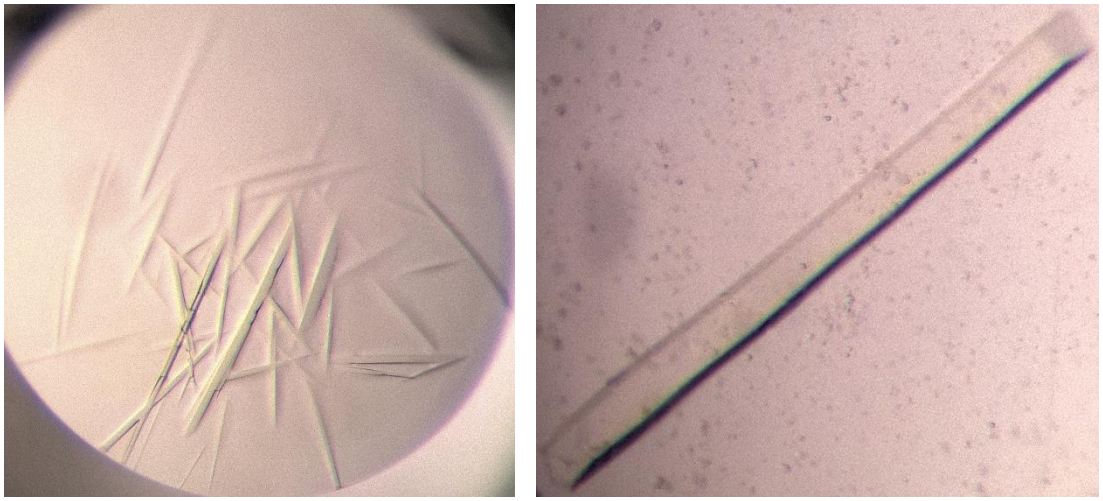


Figure 5.2 Initial Cwp19$_{27\text{-}401}$ crystallisation hits. Both formed in conditions containing 50 mM monobasic potassium phosphate. A lower PEG 8,000 concentration of 14% on the right produced a single much larger crystal (approx. 1.2 mm) than the higher concentration of 20% on the left.

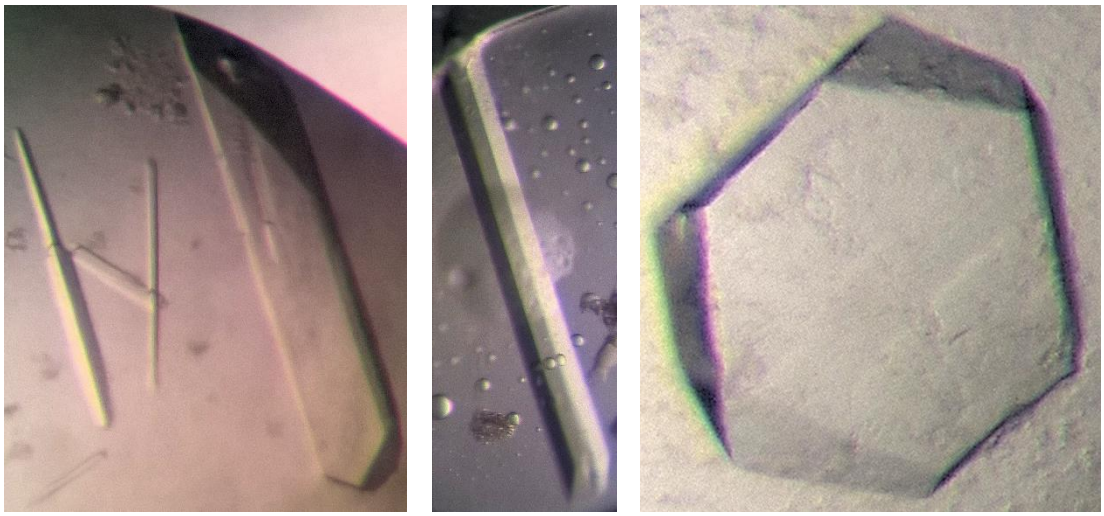


Figure 5.3 Optimised Cwp19$_{27\text{-}401}$ crystallisation hits. All optimised crystals grew considerably more quickly than the initial hits in conditions that contained carbohydrates. A change of space group from primitive monoclinic $P2_1$ to primitive orthorhombic $P2_12_12_1$ was observed for some but not others. No pattern was observed between the changes in space group and morphology.

conditions were produced using a Protein BioSolutions OptiMatrix Maker and additive screens were also performed in an attempt to reduce the degree of anisotropy (figure 5.3).

Two additive conditions were identified that produced crystals that diffracted to a higher resolution with reduced anisotropy, Molecular Dimensions Morpheus (M1) condition F7 (120 mM monosaccharides, 100 mM HEPES/MOPS pH 7.5, 40% glycerol, 20% PEG 4,000) and Morpheus II (M2) condition F7 (100 mM Monosaccharides II, 100 mM BES/TEA pH 7.5, 40% pentane-1,5-diol). The crystal that diffracted to the highest resolution (figure 5.4) was obtained in a drop containing 90% (10 mM KH_2PO_4, 18 % PEG 8,000) and 10% M1 F7 mixed 1:1 with protein at 40 mg ml^{-1}. These conditions resulted in a change of space group and cell dimension from the primitive monoclinic cell observed for H&L H11 to a primitive orthorhombic cell.

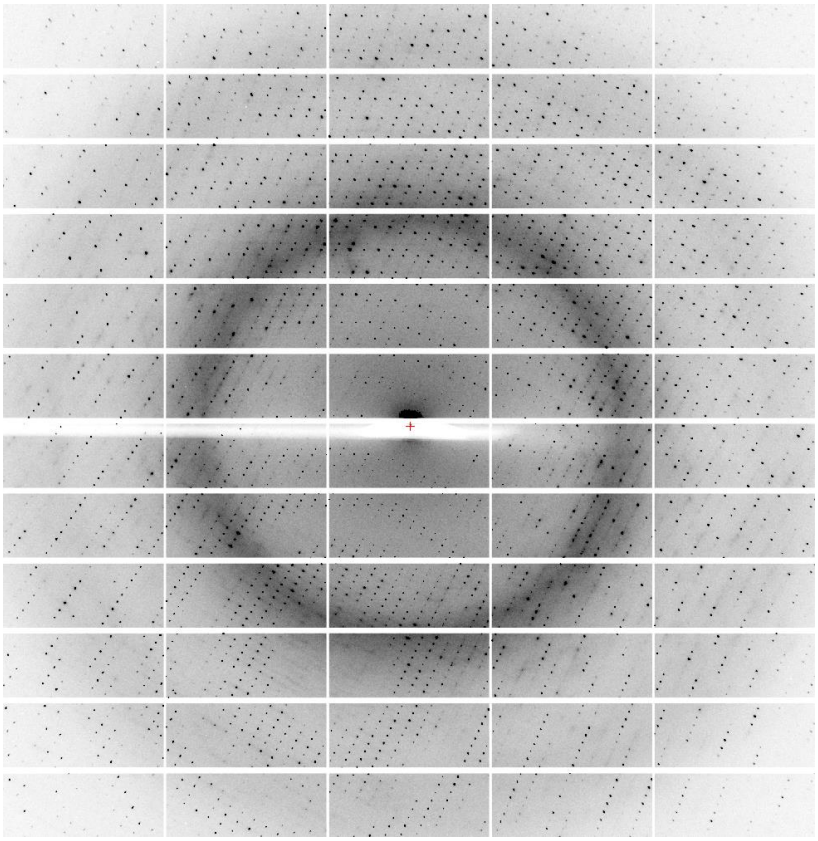


Figure 5.4 Example high resolution Cwp19$_{27\text{-}401}$ diffraction. This test shot was taken with edge of the detector positioned to capture reflections to a resolution of 2.0 Å. As spots are clearly visible into the corners of the image (approximately 1.5 Å), it is evident from the test shot that the crystal diffracts to a much higher resolution. Ultimately, a fast, low resolution sweep was collected at 2.0 Å and a slower, high resolution sweep was collected at 1.1 Å.

It was observed from other attempts at optimisation that the presence of carbohydrates seemed to reduce anisotropy and aid in crystal growth. Because of this, a significant number of screens were set up using conditions based upon earlier hits with a variety of different carbohydrates present. However, no sugars were seen to be bound in a large number of datasets collected from the crystals that these screens yielded.

Attempts at molecular replacement using these data still failed so a selenomethionine derivative was expressed according to the protocol given in chapter 3, and purified in the same way as the native protein. Crystals were obtained in drops containing 90% H&L H11 with 10% M2 F7 mixed 1:2 (protein:reservoir) with protein at 53 mg ml^{-1}. As with $Cwp84_{33-497\_C116A}$, the presence of an anomalous signal from the selenium atoms was confirmed with a fluorescence scan, the results from *CHOOCH* (Evans & Pettifer, 2001) are given in figure 5.5.

Although data were collected for MAD, the structure was ultimately determined by SAD using the CRANK2 SAD pipeline. Datasets containing 9,999 images each with oscillation angles of 0.1° for a total of 2,999.7° of data (175 GB) collected from two crystals were used to determine the structure. CRANK2 is able to calculate theoretical anomalous scattering

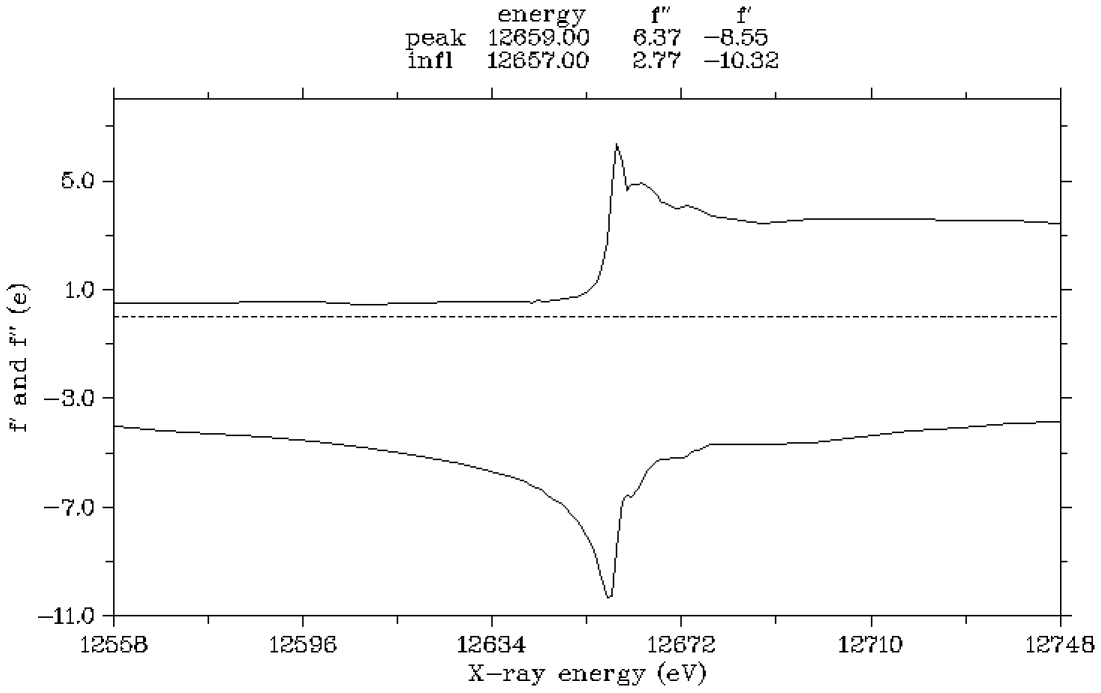


Figure 5.5 Fluorescence scan *CHOOCH* output. *CHOOCH* confirms the presence of an anomalous signal at the Se-K edge and calculates the anomalous scattering factors for structure determination.

factors. These anomalous scattering factors differed significantly from those determined by *CHOOCH*. The program was therefore run with the theoretical scattering factors and the average of the scattering factors from the fluorescence scans performed on each of the two crystals. As with $Cwp84_{33\text{-}497\_C116A}$, the number of cycles or runs of each part of the pipeline had to be increased significantly to achieve a correct solution.

The Se-SAD structure was refined and used as a model for molecular replacement with the high resolution data. Due to anisotropy, the resolution latter was cut to 1.35 Å. Crystallographic statistics are given in table 5.6.

5.3.3 The structure of $Cwp19_{27\text{-}401}$

The structure of $Cwp19_{27\text{-}401}$ has been determined by selenium single-wavelength anomalous diffraction (Se-SAD) and to a high resolution with native data using a construct coding for residues 27-401, although electron density is only visible for residues 28-388 across the two structures presented here. This construct does not contain the signal peptide, which is predicted to be cleaved between residues 24 and 25 (Petersen *et al.*, 2011), or the three C-terminal cell wall binding domains, the first of which is predicted to start at residue 402 (Finn *et al.*, 2016).

The Se-SAD structure has been determined to a resolution of 2.3 Å and contains two protein chains in the asymmetric unit with two phosphate ions, a PEG molecule and 136 water molecules, while the high resolution native structure has been determined to 1.35 Å with one protein chain in the asymmetric unit, two PEG molecules, a chloride ion and 385 water molecules. The two Se-SAD chains superpose on the high resolution structure with RMSDs of 0.27 Å (2527 atoms) and 0.28 Å (2435 atoms), while they superpose on each other with an RMSD of 0.28 Å (2408 atoms).

As predicted, $Cwp19_{27\text{-}401}$ assumes a typical TIM barrel fold, forming an eight-stranded parallel β-barrel surrounded by eight α-helices (Figure 5.7). This structure is formed by a repeating βα motif. The TIM barrel is formed by residues 33-388, it is assumed that residues 389 to approximately 401 form a disordered loop linking the TIM barrel to the first cell wall binding domain. Loops following α-helices and preceding β-strands (αβ loops) on one side of the barrel are considerably shorter than those following strands and preceding α-helices (βα loops) on the other. Longer βα

Table 5.6 Cwp19$_{27-401}$ crystallographic statistics. Inner shell statistics are given in square brackets, overall statistics are un-bracketed and outer shell statistics are given in round brackets.

	Se-SAD	High resolution native
Crystallographic statistics		
Space group	$P2_1$	$P2_12_12_1$
Cell dimensions (Å, °)	55.3, 60.4, 105.1 90, 94.2, 90	62.0, 65.6, 104.0 90, 90, 90
Resolution (Å)	[55.13-8.91] (2.38-2.30)	[65.64-7.39] (1.37-1.35)
R_{merge}	[0.142] 0.255 (0.591)	[0.067] 0.118 (0.674)
R_{meas}	[0.145] 0.260 (0.603)	[0.069] 0.123 (0.704)
R_{pim}	[0.028] 0.050 (0.116)	[0.019] 0.033 (0.200)
$CC_{1/2}$	[0.998] 0.999 (0.980)	[0.998] 0.999 (0.982)
Mean <I/σI>	[68.5] 28.2 (9.9)	[40.4] 15.4 (3.7)
Completeness (%)	[99.7] 100.0 (100.0)	[99.8] 99.5 (100.0)
Total number of reflections	[26,570] 1,631,844 (162,160)	[14,762] 2,376,913 (107,557)
Total number of unique reflections	[573] 30,986 (3,026)	[679] 93,318 (4,543)
Multiplicity	[46.4] 52.7 (53.6)	[21.7] 25.5 (23.7)
Anomalous completeness (%)	[99.8] 100.0 (100.0)	[100.0] 99.4 (100.0)
Anomalous multiplicity	[26.0] 26.6 (26.9)	[13.1] 13.3 (12.0)
CC_{anom}	[0.620] 0.448 (0.203)	[-0.484] -0.307 (-0.195)
Anisotropic delta B	9.07	11.58
Anisotropic $CC_{1/2}$ = 0.3 (Å)	1.95, 1.78, 1.53	1.04, 1.32, 0.99
Refinement statistics		
R_{work}/R_{free}	0.192/0.254	0.149/0.174
RMSDs **Bond Lengths (Å)** **Bond Angles (°)**	 0.009 1.312	 0.013 1.535
Ramachandran Statistics (%) **Favoured** **Allowed** **Outliers**	 94.8 4.5 0.7	 97.2 2.8 0
Average B-factors (Å^2) **Protein** **Ligand** **Water**	 28.8 46.6 21.2	 15.0 25.8 28.1
Number of atoms **Protein** **Ligand** **Water**	 5788 14 136	 2940 12 385
PDB Code	5OQ2	5OQ3

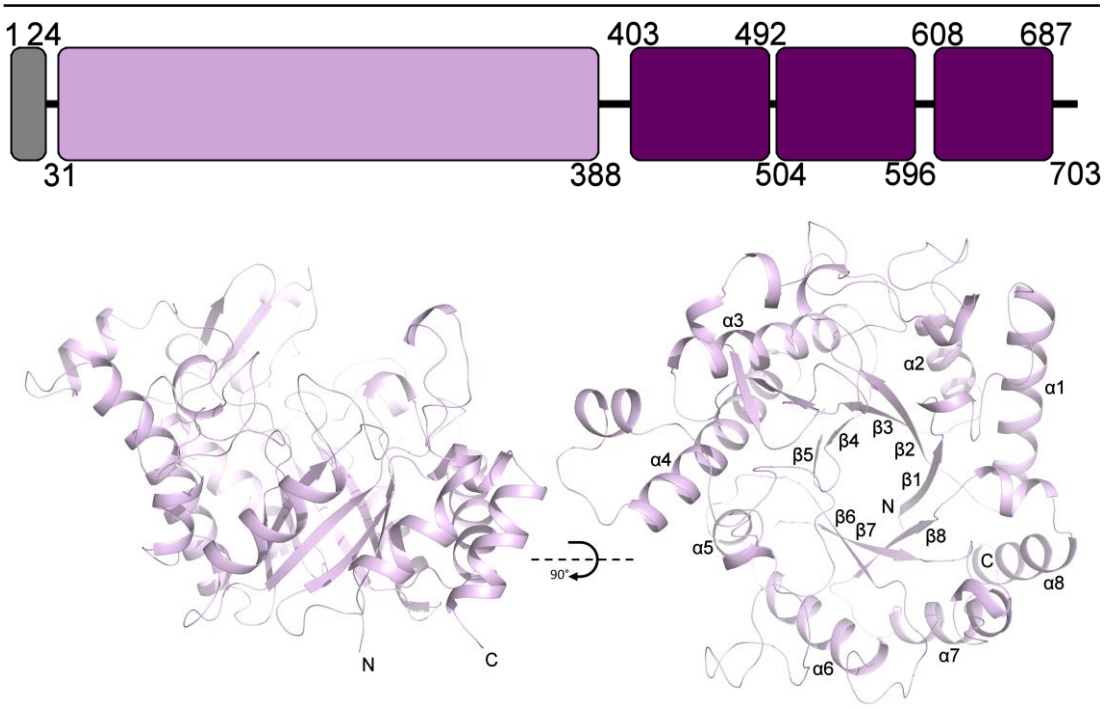


Figure 5.7 The Structure of Cwp19$_{30-388}$. A domain representation is given for full length Cwp19, the signal peptide is shown in grey, the GHL10 domain in lilac and the cell wall binding domains in purple. The present construct codes for residues 27-401, while residues 28-388, which form the GHL10 domain, are visible in the structures. This domain assumes a TIM barrel fold with an eight-stranded β-barrel surrounded by eight α-helices. The active site is formed over the centre of the barrel near the C-termini of the β-strands. The two ribbon diagrams are related by a 90 ° rotation on the x-axis.

loops than αβ loops is a common feature of TIM barrels. αβ loops frequently have the purely structural role of barrel formation, while βα loops show a significant amount of variation and form any functional sites on one side of the barrel (Wierenga, 2001).

5.3.4 Identification of the active site

Docking of simple carbohydrates to the high resolution structure of Cwp19$_{27-401}$ using SwissDock (Grosdidier et al., 2011) gave around 1250 potential modes of substrate binding. The majority of the docked ligands sat roughly centrally over the barrel (Figure 5.8). Although it is difficult from this exercise to determine exactly how any substrate binds, especially as the docked carbohydrates may not be actual substrates or products, it does give a strong indication that these regions correspond to the active site and extended binding site. This was further confirmed with a structural

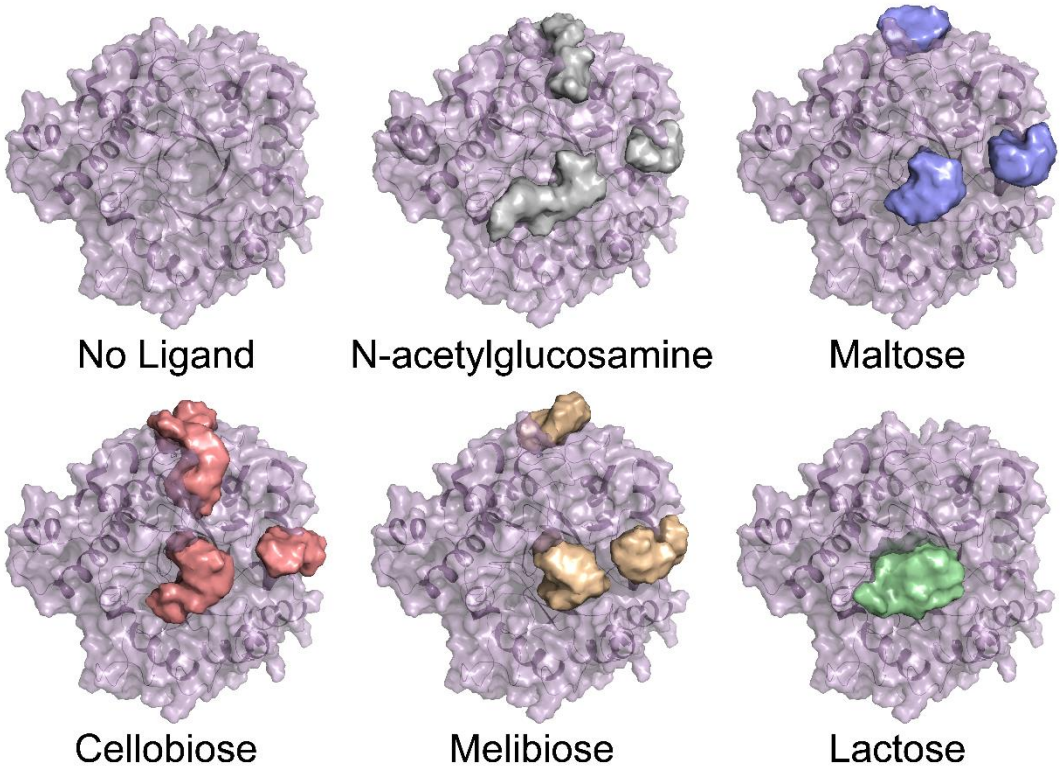


Figure 5.8 Docking results. $Cwp19_{30\text{-}388}$ is shown with the docking results for one monosaccharide and four disaccharides. Each images shows a surface representation of approximately 250 docking results for the respective carbohydrate. This produces a shape within which a sugar is predicted to bind, from this the central active site can be identified and potentially also regions to which distal portions of the substrate are able to bind. These regions are shown above and to the right of the central active site region.

alignment against $Cwp19_{27\text{-}401}$ using the DALI server (Holm & Rosenstrom, 2010), which identified hundreds of structures with significant Z-scores (Z < 2.0). PgaB (Carbohydrate esterase family 4), a subunit of a poly-β-1,6-N-acetylglucosamine deacetylase from *E. coli* (Little et al., 2014) was the closest match (Z = 23.8-26.2), followed by *Bifidobacterium bifidum* β-galactosidase (GH42, Z = 22.7) (Godoy et al., 2016) and *Solanum lycopersicum* β-mannanase 4a (GH5, Z = 22.0) (Bourgault et al., 2005). These structures showed a conserved active site in the same location as that identified by the docking. Interestingly, the putative active site in the high resolution structure shows a small amount of strong unidentified density, a formate ion fits the density well but no formate was known to be included in the crystallisation conditions so the density was left un-interpreted.

5.3.5 Peptidoglycan hydrolase assay

It has previously been suggested that Cwp19 is capable of breaking down peptidoglycan (Peltier et al. unpublished work). This was used as a starting point for the determination of an optimum pH for $Cwp19_{27\text{-}401}$ at which further activity assays could be performed. The rate of lysis of *Micrococcus luteus* cells, measured as the change in OD_{450} of a cell suspension, due to peptidoglycan breakdown was used to assess peptidoglycan hydrolase activity. Lysozyme was used as a positive control (figure 5.9).

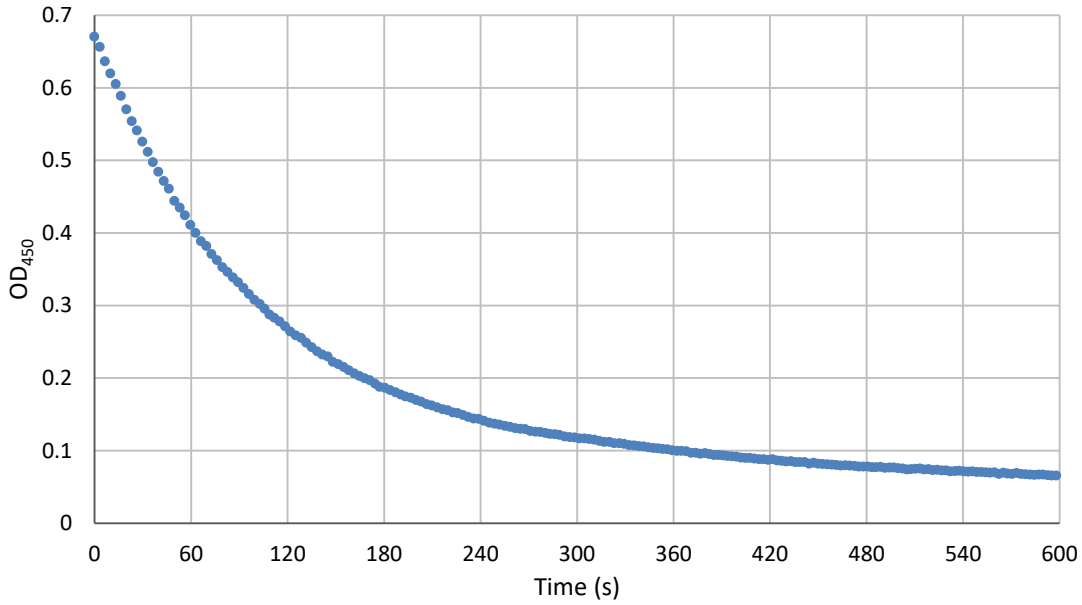


Figure 5.9. Hydrolysis of *M. luteus* cells by lysozyme. As the lysozyme breaks down the peptidoglycan forming the cell walls, the *M. luteus* cells lyse, resulting in a decrease in OD_{450}. An approximate initial rate of 0.0049 s^{-1} can be determined from the graph.

A similar profile was observed when the assay was performed with $Cwp19_{27\text{-}401}$, however, the rate of reaction was orders of magnitude slower (figure 5.10). The change in OD_{450} over two hours was measured between pHs 3.9 and 6.6. Example profiles are given in figure 5.10. Combining all results, this produced a clear bell shaped curve centred around pH 5.2-5.4 (Figure 5.11A). A faster initial rate was observed at the more acidic pHs, but after a length of time dependent on the pH, this decrease in OD_{450} stopped (Figure 5.11B). Because of this, the decrease over the first

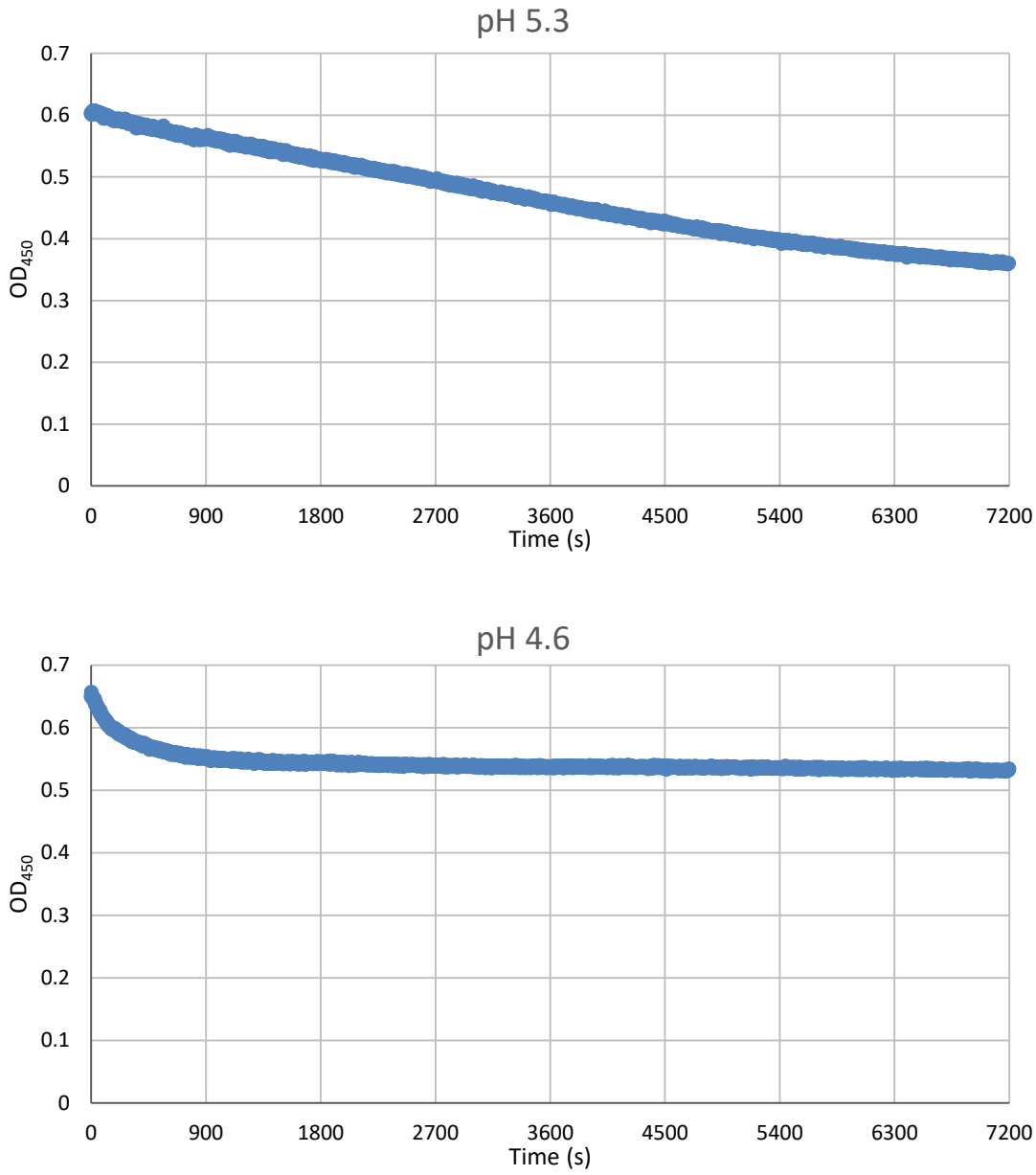


Figure 5.10 Cwp19$_{27\text{-}401}$ peptidoglycan hydrolase activity assay measurements. Example measurements are given at pH 5.3 and 4.6. Other results followed very similar profiles. It can be seen that at a relatively basic pH, the reaction proceeded as expected, while at a relatively acidic pH, although the initial rate was faster, the reaction appeared to stop after a short length of time.

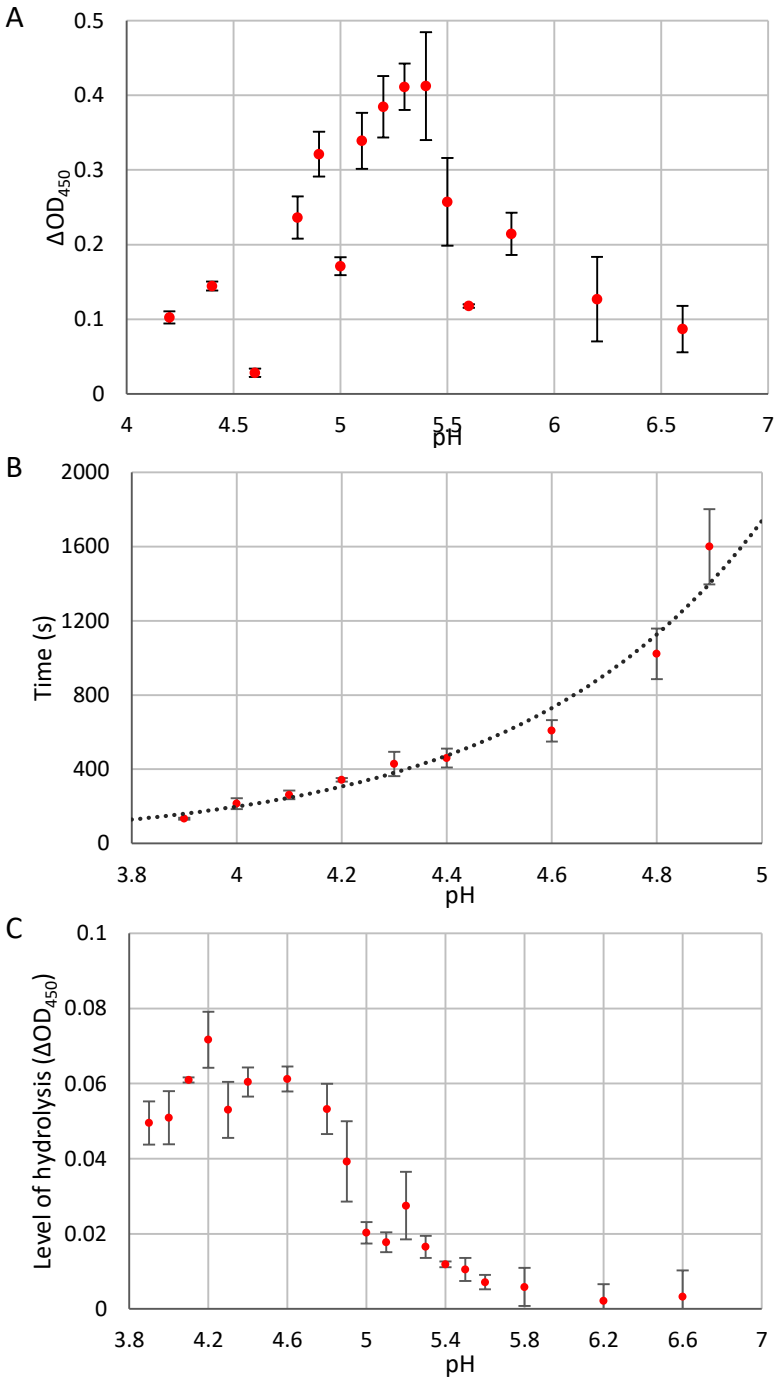


Figure 5.11 Cwp19$_{27\text{-}401}$ peptidoglycan hydrolase activity assay results. (A) ΔOD_{450} against pH over 2 hours. The greatest change in OD and therefore the greatest degree of lysis over 2 hours was observed between pH 5.2 and 5.4. **(B)** Time before the reaction stopped against pH. This closely followed an exponential pattern, strongly indicating that the cessation of the reaction was linked to pH, the precise reason for this remains unclear, but could be due to product inhibition. As the initial rate was slower at more basic pHs, it became more difficult to determine an end point, so only pHs between 3.9 and 4.9 have been included. **(C)** ΔOD_{450} against pH over 3 minutes. A considerably faster initial rate was seen at more acidic pHs, this could not be investigated beyond pH 3.9 as the cells appeared to spontaneously lyse.

5.3.6 Benedict's assay

Benedict's reagent detects the presence of reducing sugars in a sample through the reduction of soluble blue copper (II) to insoluble red copper (I) (Benedict, 1909). As the breakdown of carbohydrates is likely to increase the concentration of reducing sugars (or reducing ends of largely non-reducing molecules) in a sample, it was decided that this would be an effective method to assess the ability of $Cwp19_{27\text{-}401}$ to break down a series of potential substrates. Before the assay was performed, a range of concentrations of glucose and maltose were used to demonstrate that it is possible to observe a difference in reducing power between a monosaccharide and a disaccharide (figure 5.12).

The breakdown of starch by amylase was used as a positive control. Samples of 0.25% starch that had been incubated at 37 °C with amylase for 15 minutes had significantly lower A_{320}s than samples without amylase ($p < 0.001$, Student's T-test) (figure 5.13).

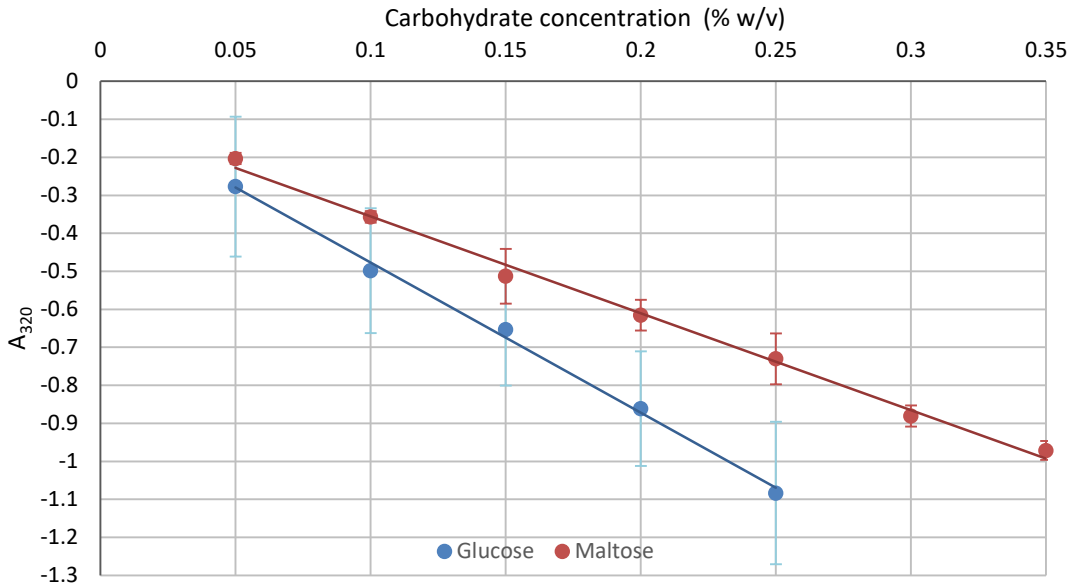


Figure 5.12 Benedict's assay calibration. Benedict's test was performed on glucose and maltose at a range of concentrations. The A_{320} of samples was observed to decrease linearly with increasing carbohydrate concentration. This linear relationship stopped beyond an A_{320} of approximately -1 to -1.1. This clearly demonstrates that it is possible to differentiate between a disaccharide and a monosaccharide.

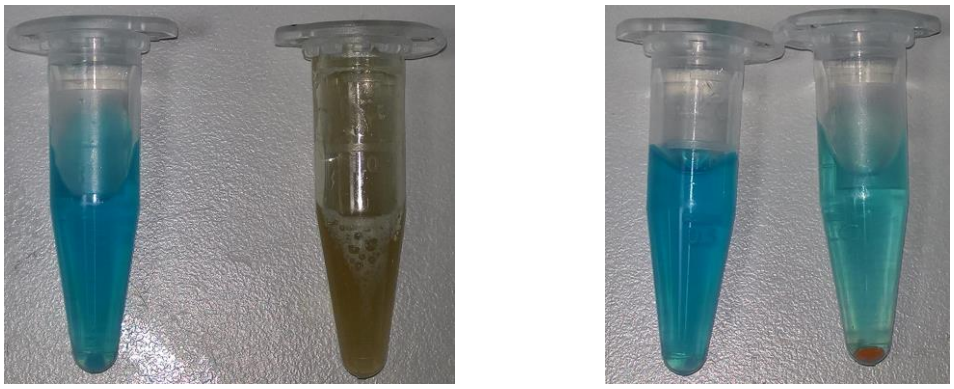


Figure 5.13 Benedict's test on starch and amylase. Before centrifugation (left) and after centrifugation (right). Benedict's test was performed on 0.25% samples of starch after incubation with and without amylase. Incubation with amylase showed a clear difference due to the formation of reducing sugars.

Table 5.14 Carbohydrates tested in Benedict's assay. All monosaccharides with unexposed reducing ends are in pyranose forms except fructose in sucrose, which is in the furanose form.

Carbohydrate	**Monosaccharides and bonds present**
Agarose	[-D-galactose-β-1,4-(3,6-anhydro-L-galactose)-α-1,3-]$_n$
Cellobiose	D-glucose-β-1,4-D-glucose
α-cyclodextrin	Cyclo-[-D-glucose-α-1,4-]$_6$
β-cyclodextrin	Cyclo-[-D-glucose-α-1,4-]$_7$
Lactose	D-galactose-β-1,4-D-glucose
Maltose	D-glucose-α-1,4-D-glucose
Melibiose	D-galactose-α-1,6-D-glucose
Pullulan	[-D-glucose-α-1,4-D-glucose-α-1,4-D-glucose-α-1,6-]$_n$
Starch	[-D-glucose-α-1,4-]$_n$ …D-glucose-α-1,6-D-glucose…
Sucrose	D-glucose-α-β-1,2-D-fructose
Trehalose	D-glucose-α-α-1,1-D-glucose

The ability of Cwp9 to hydrolyse 11 different carbohydrates was assessed (Table 5.14). None of the 11 carbohydrates tested showed significant changes in absorbance in the presence of $Cwp19_{27\text{-}401}$ ($p > 0.05$, Student's T-test).

5.4 Discussion

This work has resulted in the determination of the high resolution structure of the functional domain of Cwp19, which possess a TIM barrel fold with similarities to a wide range of other glycoside hydrolases. The diverse functions of glycoside hydrolases make it difficult to predict a function based upon the structure, but the structure can be used as a starting point for characterisation of the enzyme.

5.4.1 Active site

Probable active site residues have been identified using three methods, firstly, through docking experiments, which showed that the active site is likely to be positioned centrally over the barrel (Figure 5.8). Secondly, through comparison to the closest structural homologues identified by DALI, whose active sites are also positioned over the centre of the barrel, and finally through alignment to other proteins classified as GHL10 (Figure 5.15) and by comparison to the GHL10 HMM logo available on the Pfam website (pfam.xfam.org).

All three top DALI results identified Asp 196 as being an important residue. The equivalent in *E. coli* PgaB, Asp466, was suggested to be responsible for stabilisation of a catalytic oxazolinium intermediate (Little *et al.*, 2014) by comparison to the structures of acidic mammalian chitinase (GH18) (Sutherland *et al.*, 2011) and dispersinB (GH20) (Manuel *et al.*, 2007). While Glu161 in β-galactosidase was shown to be important to catalysis through mutagenesis and activity assays (Godoy *et al.*, 2016) and Glu204 from β-mannanase 4a was identified as part of the catalytic dyad (Bourgault *et al.*, 2005).

The other residue identified as part of the catalytic dyad in β-mannanase 4a was Glu318, which is conserved in *B. bifidum* β-galactosidase as Glu320 and was similarly shown to be important through mutagenesis and activity assays. This residue is found

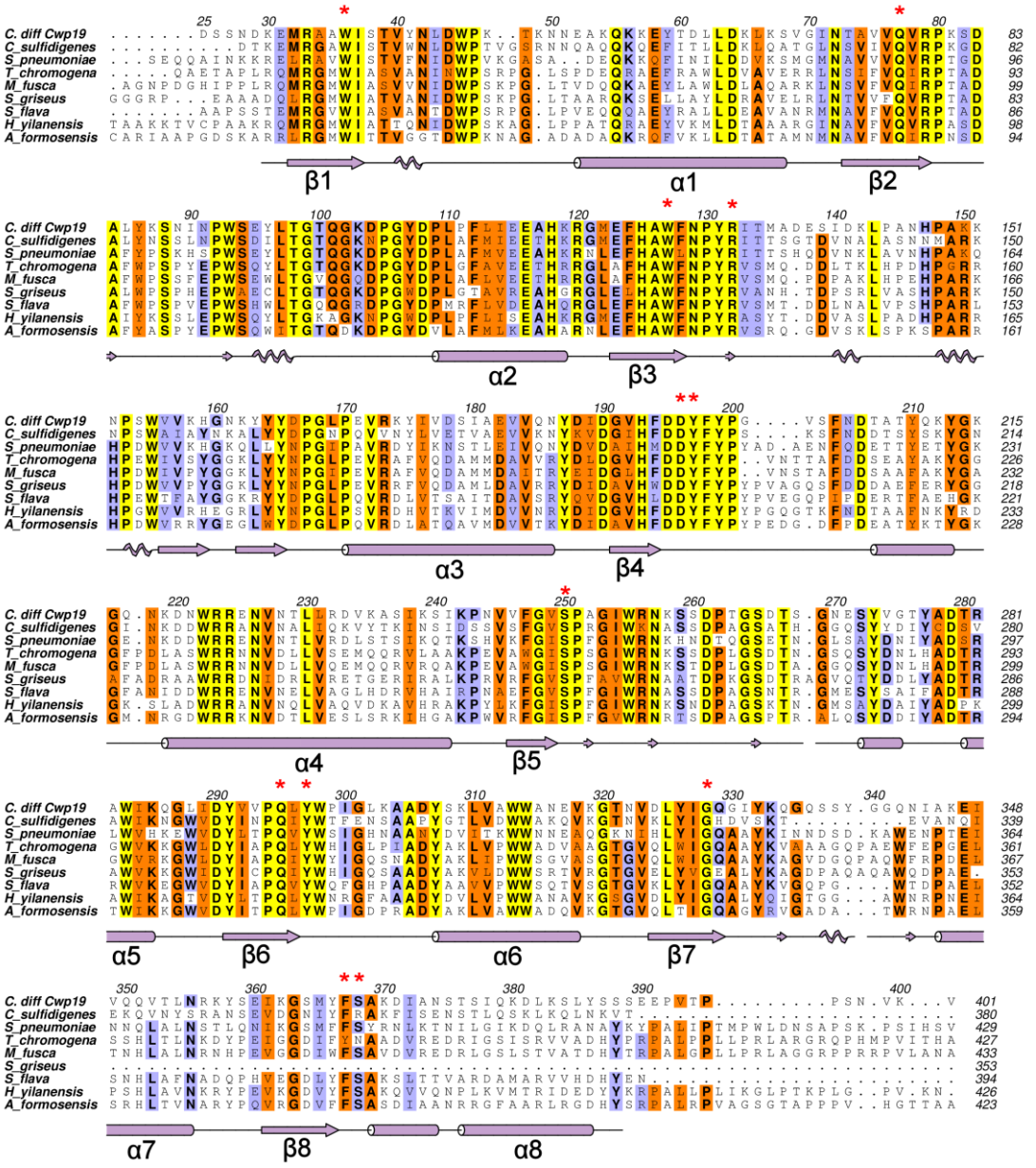


Figure 5.15 Multiple sequence alignment of Cwp19$_{25\text{-}401}$ against several BLAST results. Residues conserved across all sequences are highlighted in yellow, well conserved residues in orange, moderately conserved residues in blue. Cwp19 residue numbering is given above the alignment and secondary structure is given below. Active site residues are indicated with asterisks. The eight β-strands and α-helices that form the TIM barrel are indicated, other α-helices and β-strands are not labelled, 3_{10} helices and β-bridges are also indicated with zig-zag patterns and small arrows respectively. Species and NCBI references are as follows: *Clostridium sulfidigenes*, WP_035134795.1; *Streptococcus pneumoniae*, COD97312.1; *Thermomonospora chromogena*, SDQ39562.1; *Microtetraspora fusca*, WP_066947673.1; *Streptomyces griseus*, WP_030738522.1; *Saccharopolyspora flava*, SFT07258.1; *Herbidospora yilanensis*, WP_062357460.1; *Actinomadura formosensis*, WP_067802682.1.

at the C-terminus of β7. In $Cwp19_{27\text{-}401}$, the strand is tilted away from the centre of the barrel and the residue is replaced with Gly328, which is conserved in GHL10. This results in a significantly different shape in this portion of the active site pocket of Cwp19. As this residue is part of the catalytic dyad in β-mannanase 4a and cannot be in Cwp19, it is likely that the substrate will be orientated somewhat differently in the active site of Cwp19, interacting with different catalytic residues.

Tyr645, which is found at the C-terminus of β8, was also identified in PgaB as being important to carbohydrate binding (Little *et al.*, 2014). β-mannanase 4a was noted as possessing a cis-peptide bond between the equivalent residue, Trp360, and Glu361 that was deduced to be important for the formation of the S_1 pocket, the portion of the protein that binds the monosaccharide residue immediately before the scissile bond (Bourgault *et al.*, 2005). This aromatic residue followed by a cis-peptide is also seen in the other two DALI hits and Cwp19 as well: between Phe367 and Ser368. This cis-peptide was also observed in the structures of *Triticum aestivum* xylanase (GH18) (Payan *et al.*, 2003) and *Canavalia ensiformis* chitinase (GH18), and was noted as a "common characteristic of chitin binding proteins of family 18" that is likely to play a role in substrate binding (Hennig *et al.*, 1995). It is therefore probable that this cis-peptide is also involved in the formation of the S_1 pocket in Cwp19.

Another residue determined to be important in PgaB was Tyr432 which is found within the long β3-α3 loop. Cwp19 contains a similar extended loop, however it assumes a very different conformation. The position assumed by the side chain of Tyr432 in PgaB is, however, approximately replicated by Tyr197 in Cwp19, shortly following β4. This is adjacent to Asp196. The region surrounding these two residues shows a significant level of conservation (figure 5.15). Remaining portions of the binding site identified in PgaB are formed by loops β1-α1 and β2-α2, both of which assume different conformations in Cwp19.

The mutagenesis and activity assays on β-galactosidase also identified Asn160, Tyr289, and His371 as important active site residues (Godoy *et al.*, 2016). Asn160 is conserved in GHL10 as Asp195 in Cwp19, although as noted for PgaB, the side chain of Asp195 is largely buried. Tyr289 is conserved in GHL10 proteins (Tyr297 in Cwp19)

at the C-terminus of β6. His371 is near the centre of the β8-α8 loop, which in Cwp19 is replaced by a short helix that isn't conserved in GHL10 and has no equivalent position.

As well as the residues identified through inspection of DALI results (Asp196, Tyr197, Tyr297, Gly328, Phe367 and Ser368), the alignment of Cwp19 to other GHL10 sequences also allows the identification of Trp36, Gln77, Trp127, Arg132, Ser250, and Gln295 as conserved residues that are likely to be important to the formation of the active site and therefore substrate binding and/or catalysis (figure 5.16).

5.4.2 Other sites highlighted based on docking study

As well as docking a large number of molecules to the putative active site, SwissDock also docked a significant number of molecules to two more peripheral regions. One of these regions is formed by loops β1-α1 and β8-α8, while the other is formed by

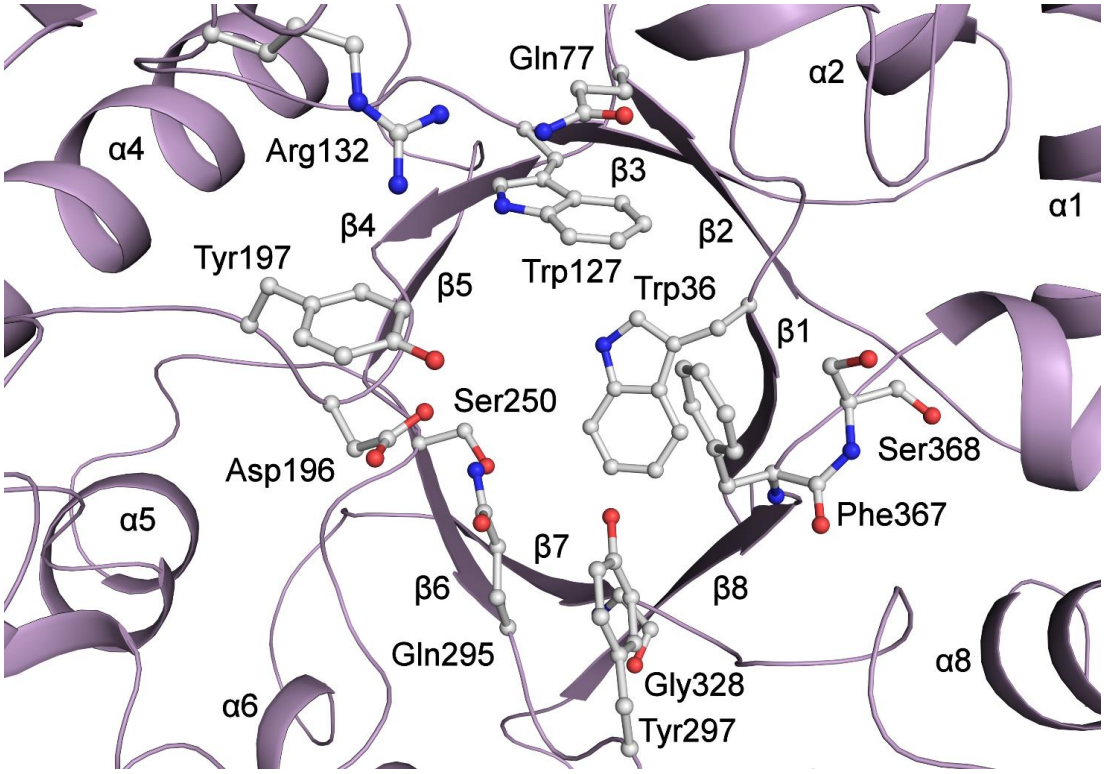


Figure 5.16 Cwp19 active site. Residues identified based upon similarities to DALI hits and other conserved residues that form part of the putative active site pocket are shown. Notably, Asp196 is likely to either be involved in intermediate stabilisation or be part of the catalytic dyad, Gly328 replaces a catalytic glutamate, and the cis-peptide bond formed by Phe367 and Ser368 is likely to be involved in formation of the S_1 pocket.

loops β2-α2 and β3-α3. These loops show significant levels of variation in GHL10 proteins and are not conserved in the closest DALI hits. The HMM logo also shows the possibility of insertions in approximately these locations. It therefore stands to reason that these regions may be responsible for substrate specificity, allowing GHL10s to cleave a range of substrates. The cis-peptide between Phe367 and Ser268, which is potentially involved in forming the S_1 pocket, forms part of the connection between the active site and the β1-α1 β8-α8 groove. It is therefore possible that the portion of the substrate before the scissile glycosidic bond sits in the β1-α1 β8-α8 groove while the portion after the scissile bond sits in the β2-α2 β3-α3 groove.

5.4.3 Activity measurements

$Cwp19_{27\text{-}401}$ is able to cleave peptidoglycan at pHs between 3.9 and 6.6, with a maximum amount of product formed at approximately pH 5.2-5.4. A faster rate was initially observed at more acidic pHs but it was not sustained. This effect clearly followed an exponential pattern strongly linking the time of onset of the arrest in activity to pH. Interestingly, the slight decrease in OD_{450} that was observed at all pHs in the controls was not observed after the reaction had stopped. Interestingly, a constant slight decrease in OD_{450} was observed at all pHs in the controls but this decrease was not observed in test samples after the reaction had stopped. Could it be possible that $Cwp19_{27\text{-}401}$ actually prevented lysis after the reaction had stopped?

Ultimately, the patterns observed here are complex and not enough information is available to fully explain them or to definitively determine an optimum pH. However, the fact that Cwp19 is able to cause lysis of the cells appears to be clear. This is very likely to be due to the hydrolysis of peptidoglycan by $Cwp19_{27\text{-}401}$. However, it is not clear which of the two glycosidic bonds in peptidoglycan that Cwp19 is capable of breaking down – N-acetylglucosamine-β-(1,4)-N-acetylmuramic acid or N-acetylmuramic acid-β-(1,4)-N-acetylglucosamine. As a significant amount of cells were lysed at pH 5.3, this pH was used for further assays.

It should be noted that even the fastest initial rate observed for $Cwp19_{27\text{-}401}$ at pH 4.2 was approximately 100 times slower than that of lysozyme at pH 6.2, while at pH 5.3 the rate was approximately 400 times slower. This indicates that Cwp19 is unlikely to

have a primary role of simply breaking down gram-positive peptidoglycan. *C. difficile* has been shown to possess an unusual form of peptidoglycan (Peltier *et al.*, 2011), so it is possible that the protein may act upon the bacterium's own peptidoglycan in a variety of possible roles.

The ability of Cwp19 to break down a range of other carbohydrate substrates was also considered. Due to the observed slow breakdown of peptidoglycan, this assay was run for four hours. It should be noted that this assay was not intended to determine a primary substrate for Cwp19, but rather to identify glycosidic linkages that Cwp19 is capable of breaking down, with an intention that this would act as a guide towards determination of the actual substrate. A range of monosaccharide residues and glycosidic linkages were tested but no statistically significant results were observed. This suggests either that Cwp19 may act on a substrate or substrates not tested in this study or that the rate of reaction for the substrates tested was too slow for a reaction to be observed.

5.4.4 PXXP motif

Glycoside hydrolase-like family 10 proteins possess a well conserved PXXP motif immediately preceding α3 – $PGLP^{170}$ in Cwp19. Notably, SH3 domains, two of which are found in Cwp14 (Sebaihia *et al.*, 2006; Monot *et al.*, 2011; Fagan & Fairweather, 2014), bind PXXP motifs (Weng *et al.*, 1995; Mayer, 2001). It is therefore, possible that there may be an interaction between Cwp19 and Cwp14. The structure of Cwp19, however, reveals that this motif is largely occluded by the beginning of α4 and the loop preceding it, particularly a short helix contained within the loop. A portion of the loop does possess slightly elevated B-factors, but it is unlikely that it will be flexible enough to facilitate binding of Cwp14. The loop is, however, poorly conserved, so it is possible that the PXXP motif in other GHL10 proteins may bind to SH3 domains.

5.4.5 Conclusions

The gene coding for Cwp19 is found within the AP locus, which is likely to be responsible for the formation of PSII. PSII has been shown to anchor the S-layer to the surface of *C. difficile* cells (Willing *et al.*, 2015; Chu *et al.*, 2016). It is therefore

possible that Cwp19 has a role in the synthesis of PSII. The recent study on Brazilian *C. difficile* strains determined that Cwp19 is one of the most abundant proteins in the S-layer (Ferreira *et al.*, 2017). This alone serves as justification for characterisation of Cwp19 to improve the very limited knowledge of the protein.

The structure of the glycoside hydrolase domain of Cwp19 consisting of residues 28-388 has been determined to a resolution of 1.35 Å. Attempts were made to crystallise $Cwp19_{27\text{-}401}$ with a number of carbohydrates, although many crystals were observed, none of the structures determined had anything bound. A number of residues that are likely to form the active site have been identified and initial characterisation of the activity of $Cwp19_{27\text{-}401}$ has been performed. $Cwp19_{27\text{-}401}$ is capable of breaking down peptidoglycan, although orders of magnitude slower than lysozyme. The enzyme either has an immeasurably slow rate of reaction for all carbohydrates tested or demonstrates a high degree of substrate specificity.

Additional work is required to further characterise Cwp19 and other proteins coded for by the AP locus so that the mechanism of PSII synthesis can be determined, which may lead to novel methods of disruption of the S-layer. Although attempts to knockout *cwp19* were previously unsuccessful (Kirby, 2011), more thorough attempts should be made, if these are still unsuccessful, this may suggest a vital role for Cwp19. Assays to determine activity on a much larger range of substrates are currently being planned. Once a substrate has been identified that Cwp19 shows a higher degree of activity against, mutagenesis can be used to confirm the role of residues identified here. This may be aided by further structural analysis.

Cwp19 has previously been classified as belonging to glycoside hydrolase-like family 10 (GHL10) based on sequence similarity (Naumoff, 2011a). This is the first structure and characterisation of a GHL10 protein and adds to the growing knowledge of glycoside hydrolases in general. More thorough characterisation of GHL10 proteins is required before a normal GH number can be assigned to the family (Bernard Henrissat, Aix-Marseille University, personal communication).

Chapter 6

Cwp2

6.1 Introduction

Cwp2 is approximately 300 residues long and possesses C-terminal cell wall binding domains. The protein has been shown to be expressed during normal growth (Calabi *et al.*, 2001), presented on the surface of the cell (Wright *et al.*, 2005) and has also been shown to form part of the spore coat (Lawley *et al.*, 2009). The functional region of Cwp2 bears similarity to very few known proteins (Altschul *et al.*, 1990), as such, when this work started, nothing was known about its structure.

The gene coding for Cwp2 is found within the SlpA locus two genes downstream from *slpA* itself, separated by *secA2* (Calabi *et al.*, 2001). This particular portion of the SlpA locus has been noted as having a significantly higher degree of SNPs than the rest of the locus (Dingle *et al.*, 2013). The same study also noted the existence of certain strains of *C. difficile* that lack a *cwp2* gene but contain a glycosylation cluster in its place and it was suggested that the contiguous *slpA*, *secA2*, *cwp2*, LmbE-like deacetylase and *cwp66* genes have undergone horizontal transfer, with each strain possessing one of at least 12 cassettes.

The majority of patients suffering from a *C. difficile* infection (CDI) appear to produce antibodies against Cwp2 (Wright *et al.*, 2008). The fact that these patients still show symptoms suggests that antibodies against Cwp2 are not protective against CDI. This, coupled with the fact that not all strains possess a *cwp2* gene (Dingle *et al.*, 2013), suggests that Cwp2 is not essential for virulence. However, it is possible that the bacterium could counter the lack of Cwp2 in unknown ways.

cwp2 is two genes upstream of *cwp66*, the protein product of which has been shown to possess adhesive properties: after heat shock, Cwp66 expression is increased and antibodies raised against it are able to compromise cellular adhesion. However, this effect was not seen without prior heat shocking (Waligora *et al.*, 2001). It has previously been claimed that *cwp2* is at the start of a polycistronic operon also coding for the LmbE-like deacetylase and Cwp66, and it was suggested that the three proteins may form an adhesin complex (Savariau-Lacomme *et al.*, 2003). The sequencing of the *C. difficile* genome has since demonstrated the presence of a terminator between *cwp2* and *lmbE*, so *cwp2* is not part of the polycistronic *lmbE*-

cwp66 message (Baerends *et al.*, 2004; Sebaihia *et al.*, 2006; Monot *et al.*, 2011). The proposed adhesin complex is also yet to be observed.

A *cwp2* knockout strain demonstrated impaired adhesion to Caco-2 cells and showed an increase in toxin release (Kirby, 2011; Bradshaw *et al.*, 2017a). The ability to adhere to mammalian cells demonstrates a possible role in host cell adhesion, while the increase in toxin release could be due to a range of factors including the disruption of the S-layer, which leads to increased stress upon the bacterium.

The presence of Cwp2 in S-layer extracts and in spore coats, the potential role in host cell adhesion and the apparent stress placed on the bacterium when Cwp2 is not present demonstrate the importance of the protein. Structural studies of Cwp2 may shed more light on the protein and its functions.

6.2 Methods

6.2.1 Previous work

A construct coding for Cwp2 lacking the signal peptide and cell wall binding domains (residues 27-322) was previously cloned, expressed and purified by Jon Kirby (2011). Crystallisation trials were performed but no crystals were observed for an extended period of time.

6.2.2 Data collection and Processing

Crystals from the aforementioned trials were observed in JCSG condition G11 (2.0 M ammonium sulphate, 0.1 M bis-tris, pH 5.5) after 5 years and were cryo-protected by addition of ammonium sulphate to a final concentration of approximately 3 M before flash cooling in liquid nitrogen. Four datasets, each consisting of 900 images with 0.2° oscillations, were collected from four crystals on beamline I04 at Diamond Light Source. Data were indexed and integrated with *DIALS* (Waterman *et al.*, 2016) and scaled with *AIMLESS* (Evans & Murshudov, 2013). The data were processed in space group $P6_322$. A high resolution cut-off of 1.90 Å was selected based on an anisotropic $CC_{1/2}$ of 0.3. This cut-off closely coincided with a cut-off that might have been selected

based on all data at around 1.80-1.85 Å due to a relatively low anisotropic delta-B of 4.1 $Å^2$.

6.2.3 Structure solution

Cwp2 has 37% identity and 56% similarity to Cwp8. Polyalanine models based upon the three functional domains of Cwp8 (5J6Q) (Usenik *et al.*, 2017) were generated and input into Phaser (McCoy *et al.*, 2007). Once the three domains had been placed, density modification was performed with PARROT (Cowtan, 2010) and 50 cycles of model building and refinement were performed with *Buccaneer* (Cowtan, 2006) and REFMAC5 (Murshudov *et al.*, 2011). The resulting structure was manually completed with rounds of *coot* (Emsley & Cowtan, 2004) and REFMAC5. The structure was validated with MolProbity (Chen *et al.*, 2010).

6.2.4 Flexibility analysis

To analyse whether the differing orientations observed in domain 2 of Cwp2, Cwp8 and LMW SLP were physiological or crystallographic artefacts and whether the domain was able to move relative to the other domains, flexibility analysis was performed. The three structures were prepared using PyMol (www.pymol.org) and MolProbity. Normal mode eigenvectors were determined for the structures of the proteins using Elnemo (Suhre & Sanejouand, 2004) while FIRST (Jacobs *et al.*, 2001) and FRODA (Wells *et al.*, 2005) were used to determine flexibility of the proteins. Ten non-trivial modes were analysed and inspected with PyMol to identify modes of motion that resulted in greater similarities between the structures.

6.3 Results

6.3.1 Data collection

Crystals of $Cwp2_{27\text{-}322}$ were observed to have formed in JCSG condition G11 (2.0 M ammonium sulphate, 0.1 M bis-tris, pH 5.5). 900 degrees of data were collected from each of four crystals that had been cryoprotected by addition of 1 µl 3.4 M ammonium sulphate to the drop (figure 6.1).

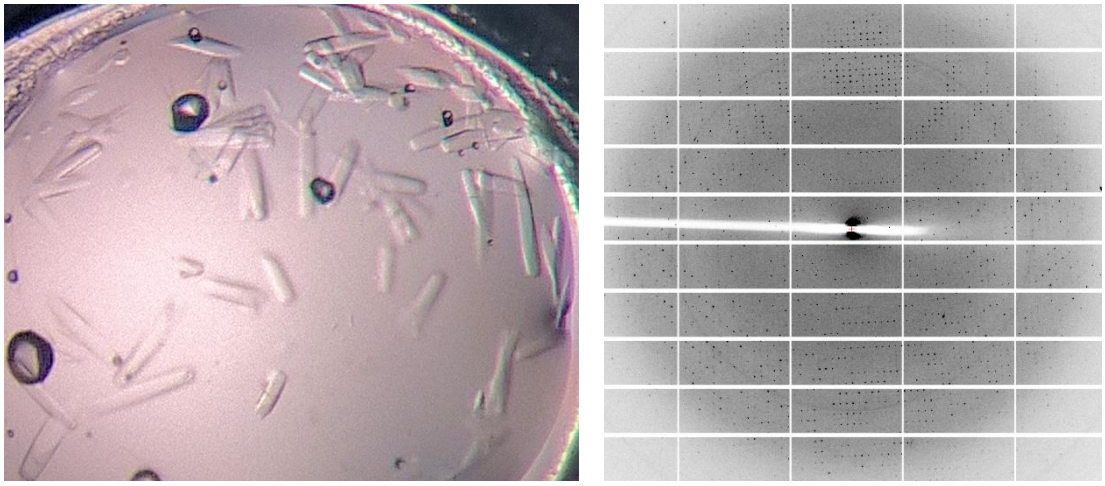


Figure 6.1 Cwp2$_{27\text{-}322}$ crystals and diffraction. Data were collected from 4 crystals, an example diffraction image is shown, reflections were clearly visible to 2.5 Å, the resolution was ultimately cut to 1.9 Å.

6.3.2 Structure solution

A PDB BLAST (Altschul *et al.*, 1990) (figure 6.2) showed low similarity of Cwp2$_{27\text{-}322}$ to any protein not from *C. difficile*, indicating a potentially new fold. Likewise, submitting the sequence of Cwp2$_{27\text{-}322}$ to Swiss Model (Guex & Peitsch, 1997) yielded only a small number of very short models with poor scores. The sequence was submitted to Robetta (Song *et al.*, 2013) and Phyre2 (Kelley *et al.*, 2015), which were able to generate models for the whole structure but molecular replacement using these models was unsuccessful. As Swiss Model and Phyre2 report the PDB codes of the structures used in generating the model, some of these were also used as models for molecular replacement with both Phaser (McCoy *et al.*, 2007) and MOLREP (Vagin

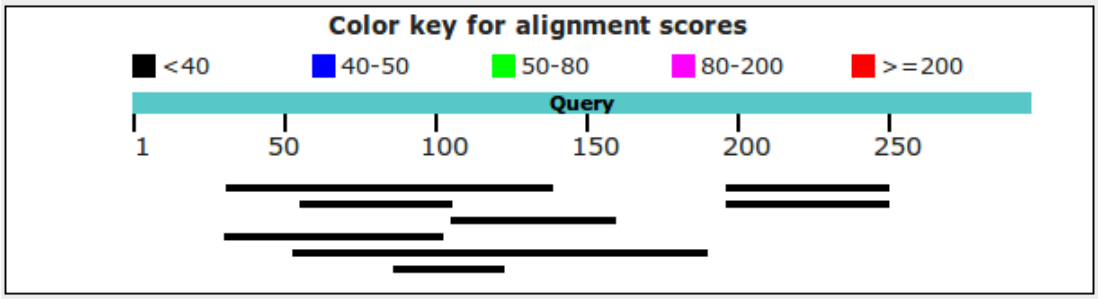


Figure 6.2 Cwp2$_{27\text{-}322}$ PDB BLAST. The functional domains of Cwp2 showed no significant similarity to any previously determined structures. This is a strong indication that molecular replacement will be difficult or even impossible.

& Teplyakov, 2010), however, these attempts did not result in a correct solution either.

Another feature of the report produced by Phyre2 is a prediction of secondary structure with associated certainties for each residue. Two extended regions were observed in this prediction of closely linked secondary structure elements with a high degree of certainty, specifically residues 147-194 and 282-320. Due to a lack of loops between the predicted secondary structure features, it was determined to be likely that they formed largely uninterrupted structural motifs within the final structure and that *ab initio* models of these regions may accurately reflect the structure of Cwp2.

Robetta fragments were generated for the two regions which were input into Rosetta (Bonneau *et al.*, 2002). 100,000 models were generated for each region and scored. The two highest scoring models (figure 6.3), one for each region, were used for molecular replacement with Phaser, however, once again, this did not yield a correct solution. The next step in this method would have been to cluster the models and produce a range of ensembles for Phaser, which increases the chances of molecular replacement being successful, particularly with relatively poor models, however this was not attempted due to the publication of the full length structure of Cwp8 (Usenik *et al.*, 2017).

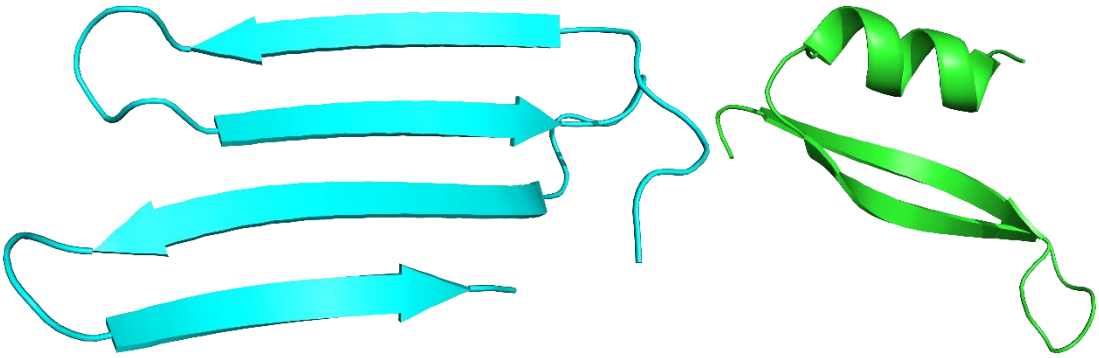


Figure 6.3 Rosetta models. The two best models produced by rosetta for the two regions of secondary structure predicted by phyre2 with the highest degree of certainty. Initial attempts at molecular replacement with these models were unsuccessful.

The functional region of Cwp8 possess an extended three domain fold. Cwp2 and Cwp8 share 27% identity and 44% similarity (Altschul *et al.*, 1990). An attempt was made to solve the structure of $Cwp2_{27\text{-}322}$ with Phaser using structure of Cwp8 with the cell wall binding domains removed as a model. Once again, this was unsuccessful. The three functional domains were also unsuccessfully input into Phaser separately. Finally, polyalanine models based on the three domains were input into Phaser. Based on cell dimensions of a = b = 134.8 Å, c = 102.8 Å, α = β = 90.0°, γ = 120.0° and a calculated molecular mass of 31.9kDa, Matthews coefficients of 4.25 for one molecule, and 2.12 for two molecules were calculated with solvent contents of 71.1% and 42.1% and probabilities of 1% and 99% respectively. This strongly indicates the presence of two molecules within the asymmetric unit. Phaser was only able to place one copy of each domain with translation function (TF) Z-scores of 8.26, 7.11 and 8.46 for domains one, two, and three, respectively. TF Z-scores below 4 are generally taken to mean that a correct solution has not been found, while TF Z-scores above 8 indicate a correct solution (Airlie McCoy, unpublished). Accurate models frequently produce TF Z-scores of 12 or more.

The TF Z-scores produced by Phaser for the three domains indicate weak, but correct, solutions. This was supported by R_{work} and R_{free} values from REFMAC5 (Murshudov *et al.*, 2011) of 0.519 and 0.528. Automated model building and refinement with *Buccaneer* (Cowtan, 2006) and REFMAC5 after density modification with PARROT (Cowtan, 2010) resulted in R_{work} and R_{free} values of 0.469 and 0.506. The structure was ultimately completed by manual building with COOT (Emsley & Cowtan, 2004) and refinement with REFMAC5 to final R_{work} and R_{free} values of 0.216 and 0.245. These figures indicate that despite the high solvent content, there was only one molecule within the asymmetric unit. This was the result of a large solvent channel on the 6_3 axis (figure 6.4).

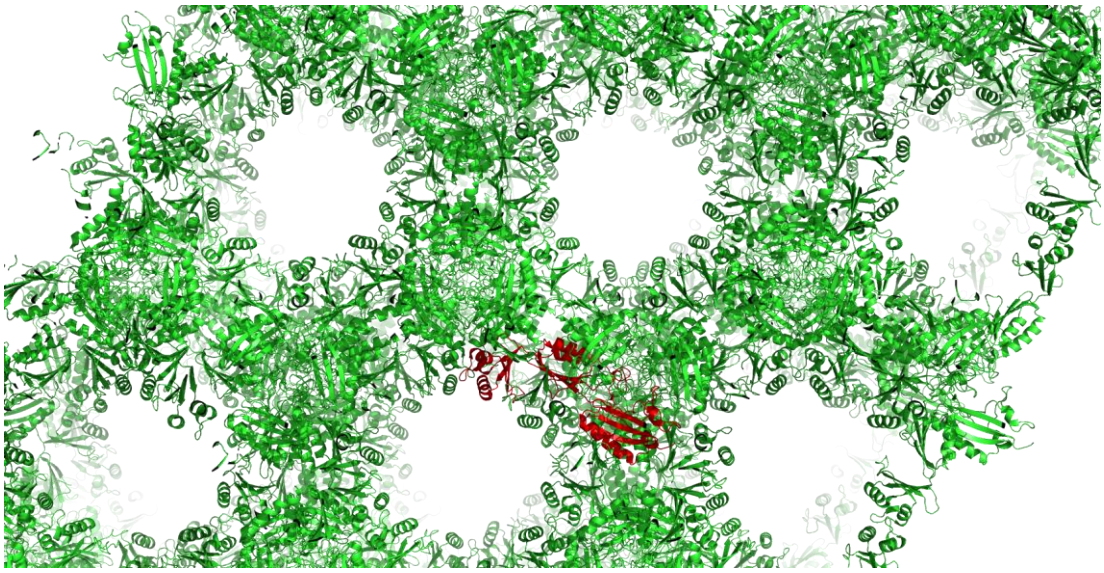


Figure 6.4 Cwp2$_{27\text{-}322}$ crystal packing. A single molecule of Cwp2$_{29\text{-}318}$ is shown in red. A large solvent channel is visible looking down the 6_3 axis. This explains the high solvent content and single molecule in the asymmetric unit. Six pairs of symmetry related helices can also be seen in the solvent channels, with three of the pairs further into the image due to the screw axis.

6.3.3 The Structure of Cwp2$_{29\text{-}318}$

The structure of the functional domains of Cwp2 (residues 29-318) has been determined to a resolution of 1.90 Å. The asymmetric unit contains one protein chain, 20 sulphate ions and 229 water molecules. Crystallographic statistics are given in table 6.5. The protein possesses a similar extended three domain fold to that of the recently determined Cwp8 (Usenik *et al.*, 2017). Despite a lack of significant sequence similarity to either protein, domains 1 and 2 of LMW SLP also possess similar folds, while the structure of domain 3 of LMW SLP, which has been shown to be responsible for the formation of the H/L complex (Fagan *et al.*, 2009), is currently unknown.

Like Cwp8, domains 1 and 3 of Cwp2 possess two-layer sandwich folds consisting of a four-stranded mixed β-sheet and two α-helices while domain 2 has a smaller antiparallel β-sheet, one α-helix and a hairpin loop (figure 6.6). All three domains show high similarity to the equivalent domains in Cwp8 and individually superpose on the structure well - with an RMSD of 3.47 Å (all atom alignment, 924 atoms - RMSDs for individual domains are given in table 6.7). In the case of domain 2 while

the β-sheet and α-helix are conserved, the loop region shows significant differences.

Table 6.5 Cwp2$_{27-322}$ crystallographic statistics. Data in the lowest resolution shell are given in square brackets, data in the highest resolution shell are given in round brackets.

Space group	$P6_322$
Cell dimensions (Å, °)	134.8, 134.8, 102.8, 90.0, 90.0, 120.0
Resolution (Å)	[120.83 – 9.11] (1.94 – 1.90)
R_{merge}	[0.051] 0.166 (4.497)
R_{meas}	[0.052] 0.169 (4.624)
R_{pim}	[0.008] 0.021 (0.772)
$CC_{1/2}$	[1.000] 1.000 (0.685)
Mean <I/σI>	[74.7] 21.6 (1.5)
Completeness (%)	[100.0] 100.0 (100.0)
Total reflections	[29,049] 2,741,248 (97,683)
Total unique reflections	[495] 43,871 (2,766)
Multiplicity	[58.7] 62.5 (35.3)
Refinement statistics	
R_{work}/R_{free}	0.215/0.245
RMSDs **Bond Lengths (Å)** **Bond Angles (°)**	 0.010 1.440
Ramachandran Statistics (%) **Favoured** **Allowed** **Outliers**	 97.2 2.4 0.4
Average B-factors **Protein** **Sulphate** **Water**	 41.6 92.5 40.3
Number of atoms **Protein** **Sulphate** **Water**	 2238 100 229
PDB Code	5NJL

The two domains superpose with an RMSD of 3.68 Å (261 atoms). However, with domains 1 and 3 superposed, domain 2 of Cwp2 is rotated approximately 40° relative to domain 2 of Cwp8.

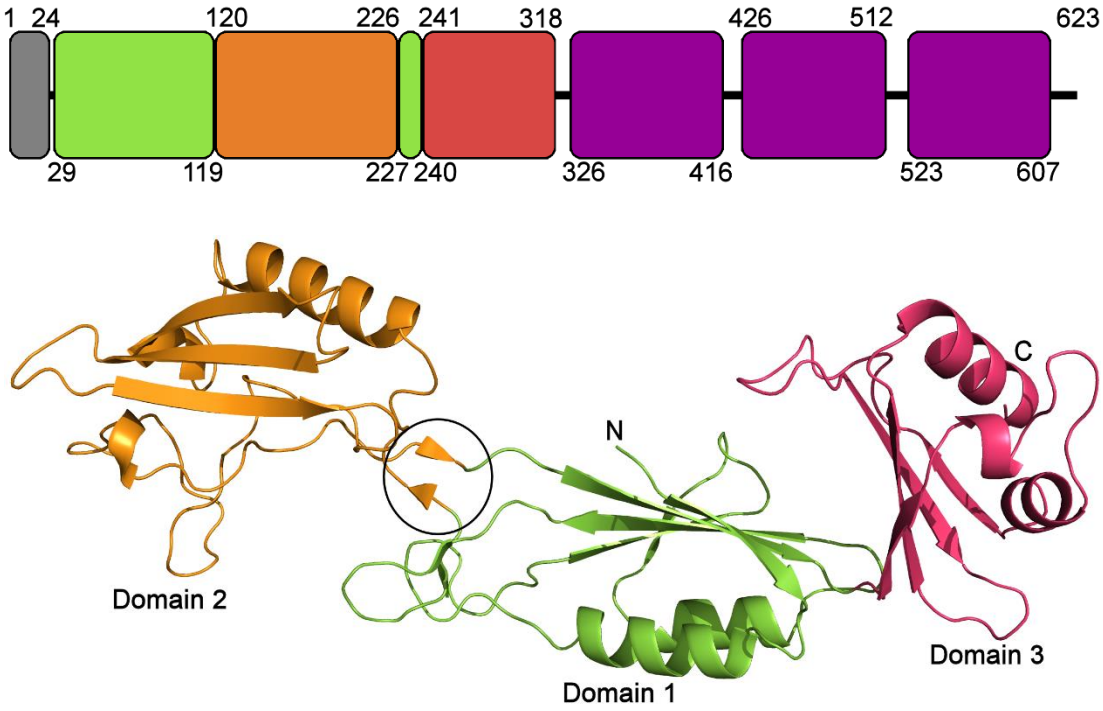


Figure 6.6 The structure of Cwp2$_{29\text{-}318}$ – Domain 1 is shown in green, domain 2 in orange, and domain 3 in pink. The functional region of Cwp2 assumes an extended three domain structure. Domain 1 possesses a two-layer sandwich fold. Domain 2 has a single α-helix, a small β-sheet and significant loop regions and is connected to domain 3 via a strand in domain 1. Domain 3 forms a similar two-layer sandwich fold to domain 1. The hinge region, which allows movement of domain 2, is circled.

Table 6.7 RMSDs for domain alignments – RMSDs are given in Å. The number of atoms in the alignment is given in brackets.

		Cwp8		**Cwp2**	
		Domain 1	**Domain 3**	**Domain 1**	**Domain 3**
LMW SLP	**Domain 1**	1.79 (263)	N/A	2.32 (271)	3.57 (96)
Cwp2	**Domain 3**	3.62 (98)	1.31 (276)	N/A	
	Domain 1	1.03 (309)	8.36 (237)		
Cwp8	**Domain 3**	7.14 (290)			

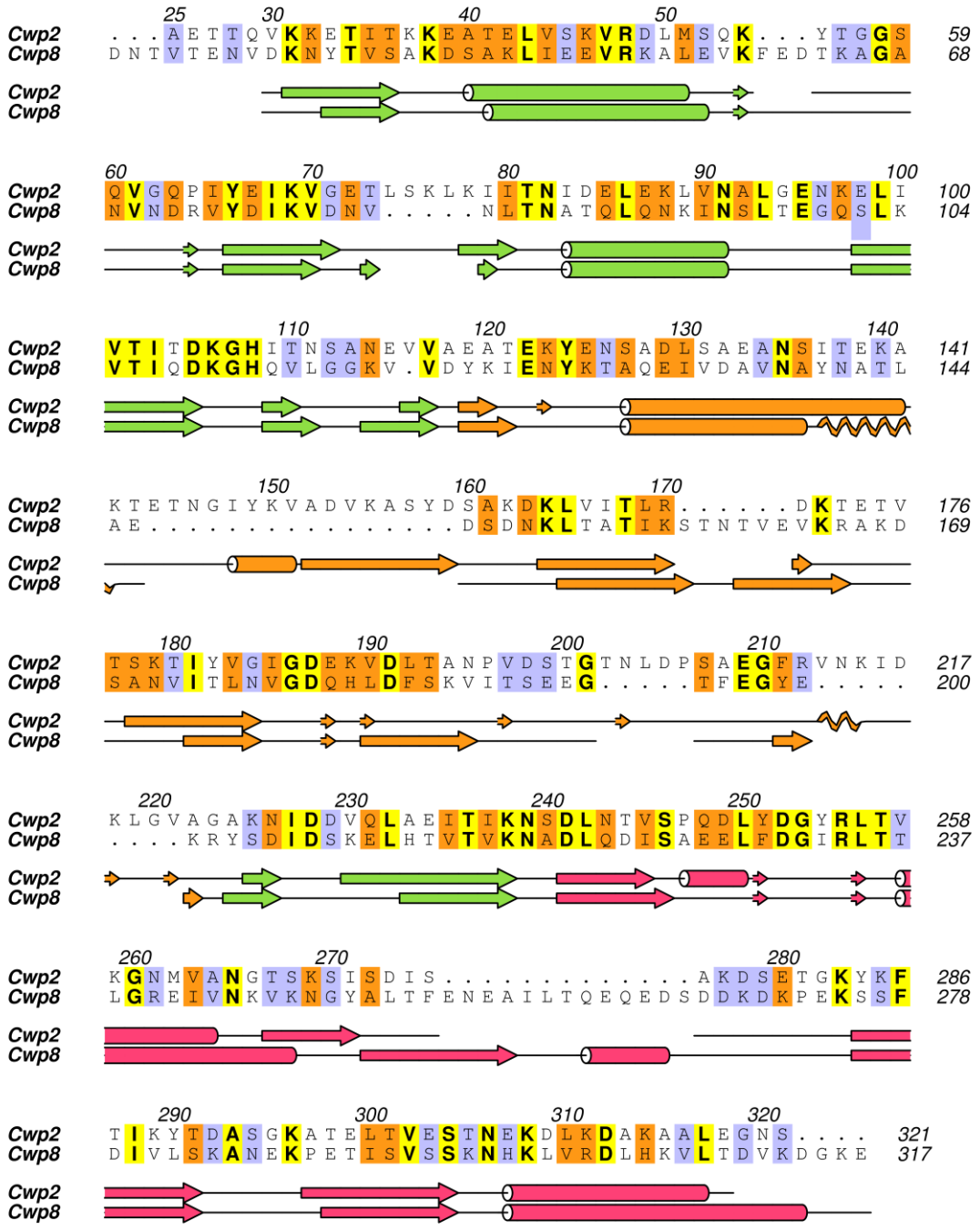


Figure 6.8 Sequence alignment of Cwp2$_{25\text{-}322}$ and Cwp8$_{36\text{-}317}$ – Both sequences are taken from strain 630. The secondary structures of the two protens are given below the sequence, domain 1 is coloured green, domain 2 in orange and domain 3 in pink. The numbers of Cwp2 residues are given above the sequence. Completely conserved residues are highlighted in yellow, moderately conserved residues in orange, and slightly conserved residues in blue. Generated with Clustal Omega (Sievers *et al.*, 2011) and Aline (Bond & Schuttelkopf, 2009).

The sequences of Cwp2 and Cwp8 are compared in figure 6.8, it can be seen that the greatest degree of variation between the two proteins is in domain 2. This domain also shows a greater degree of structural variation between Cwp2, Cwp8 and LMW SLP than domain 3 of Cwp2 and Cwp8 and domain 1 of all three proteins. Although the α-helix and β-sheet are shared by all three, the loop region shows a significantly different conformation and is considerably extended in LMW SLP.

6.3.4 Flexibility analysis

Analysis of the structures of Cwp2, Cwp8 and LMW SLP with FRODA demonstrated a significant degree of hinging between domain 1 and 2 of the three proteins. Although the hinging between domains 1 and 2 of Cwp2 and Cwp8 did allow the two domains of the respective proteins to assume orientations somewhat closer to the other protein (approximately 10° closer), the closest orientations achieved were by no means identical (figure 6.9). A much smaller degree of flexibility was observed between domains 1 and 3. This demonstrated that the orientations within the structures may be partially attributed to crystallographic artefacts and that the domains are able to move relative to each other to a degree, but this does not fully account for the differences between the three proteins.

6.4 Discussion

This work has resulted in the determination of the structure of the functional domains of Cwp2 to 1.9 Å. Work done alongside this performed by Jon Kirby demonstrated that Cwp2 shows adhesive properties *in vitro* (Bradshaw *et al.*, 2017a).

Cwp2 is constitutively expressed and found on the surface of *C. difficile* cells during normal growth (Calabi & Fairweather, 2002; Wright *et al.*, 2005) and is also found in the spore coat (Lawley *et al.*, 2009). However, the gene is not found in certain strains (Dingle *et al.*, 2013). A significant number of CDI patients raise antibodies to Cwp2, which may suggest that antibodies raised against the protein are not protective (Wright *et al.*, 2008). The presence of Cwp2 on the cell surface during normal growth does, however, suggest that it is required for some cellular process(es).

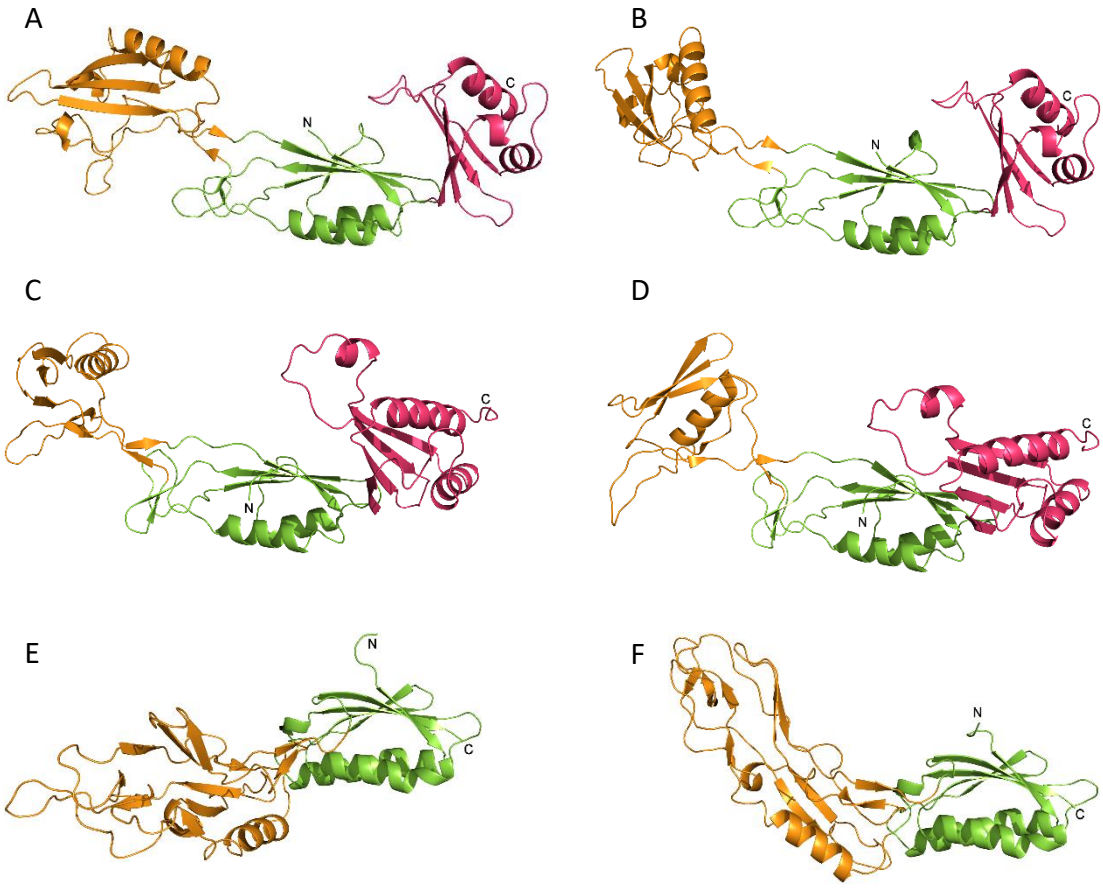


Figure 6.9 Comparison between the functional domains of Cwp2, Cwp8 and LMW SLP and flexibility analysis – (A, C, E) The structures of Cwp2, Cwp8 (5J6Q), and LMW SLP (3CVZ), respectively. Domains are coloured in the same was as figure 6.6. The flexibility of these structures and the hinge region between domains 1 and 2 was analysed with FIRST and FRODA. (B, D, F) Cwp2, Cwp8 and LMW SLP after rotation of domain 2 around the hinge region. Domain 2 of both Cwp2 and Cwp8 is able to assume an orientation closer to that seen in the other protein. Although domain 2 of LMW SLP shows a considerably higher degree of difference to domain 2 of Cwp2 and Cwp8 than the two do to each other, it is still able to rotate and assume a position somewhat closer to that seen in Cwp2 and Cwp8.

In the recently reported structure of Cwp8, to which this work has shown that Cwp2 bears significant structural similarity, Usenik *et al.* (2017) demonstrated that domain 2 is likely to form the most surface exposed region of the protein, while domains 1 and 3 are more buried. This is also likely to be true for Cwp2 due to the high similarity between the two structures. Domain 2 in both proteins consists of an α-helix, a three-stranded β-sheet and an extended loop region (Figure 6.10), although there are significant differences in the loop regions. The two domains also show significantly different orientations in their respective structures, related to each other by a rotation of approximately 40° (figure 6.9). Domain 2 of LMW SLP, bears a level of similarity to domain 2 of Cwp2 and Cwp8, although the β-sheet in this protein has a slightly different orientation relative to the helix and the loop region is considerably larger.

Domain 2 of LMW SLP has been shown to exhibit low sequence identity between strains, which is likely to be permitted by the significant number of loops within the structure (Fagan *et al.*, 2009) (figure 6.10c). The protein exhibits both immuno-evasive capabilities (Spigaglia *et al.*, 2011) and adhesive properties (Merrigan *et al.*, 2013). The variability of domain 2 has been suggested to play a role in immuno-evasion (Fagan *et al.*, 2009) while variability of the whole protein has been shown to have an effect on the strength of cell adhesion (Merrigan *et al.*, 2013). As domain 2 is very likely to be the most surface exposed part of the structure, it is a logical conclusion that this domain is likely to be primarily responsible for both immuno-evasion and cell adhesion. Similarly, Cwp2 and Cwp8 both show increased amount of SNPs compared to other proteins within the slpA locus (Dingle *et al.*, 2013). It is therefore possible that the differences in domain 2 of Cwp2 and Cwp8 may too have an effect on adhesive abilities and that the variation may be linked to immuno-evasion. However, the cross-reactivity of sera against Cwp2 (McCoubrey & Poxton, 2001) casts doubt on the possibility of a role for Cwp2 in the latter.

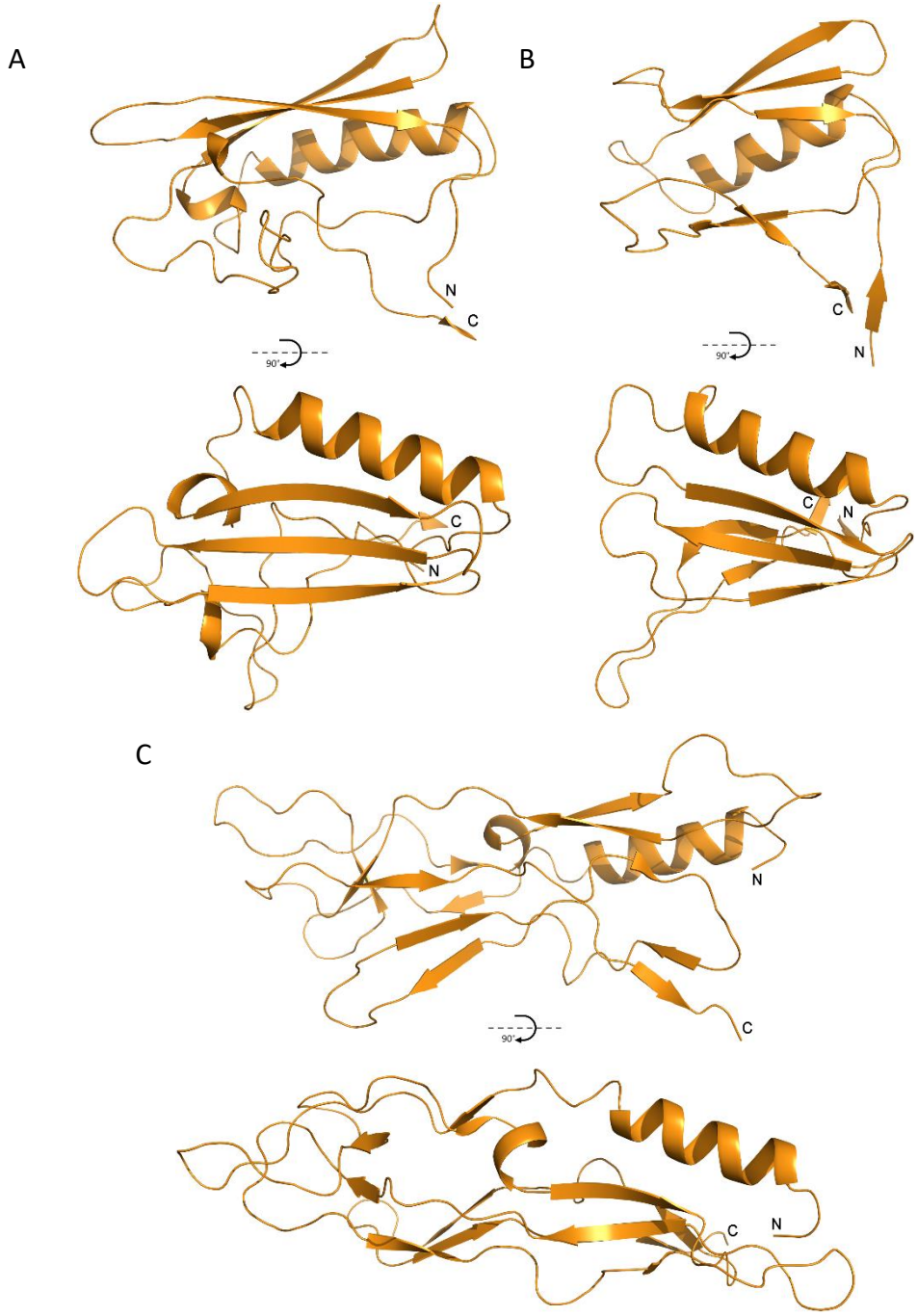


Figure 6.10 Comparison of Domain 2 of Cwp2, Cwp8 and LMW SLP - (A) Cwp2. (B) Cwp8. (C) LMW SLP. All three share a conserved α-helix and β-sheet. All three also possess extended loop regions whose structures are not conserved between the three proteins. The loop regions of LMW SLP, the variability of which has been linked to variations in adhesion, are considerably longer than those seen in Cwp2 and Cwp8.

Due to the differing orientations of domain 2 across the three structures and a degree of isolation exhibited by the domain, it is tempting to suggest that the interface between domains 1 and 2 of the three proteins may be able to act as a hinge allowing domain 2 to move relative to domains 1 and 3. Notably, this may allow domain 2 in Cwp2 to assume a position and orientation closer to that of domain 2 in Cwp8 and vice versa. Domain 2 of LMW SLP is significantly different to the other two, but may still exhibit flexibility in this hinge region.

To examine this hypothesis, the flexibility of the three structures was computationally analysed using FIRST and FRODA. By probing the structures with physiologically relevant amounts of energy, it was possible to observe a hinging movement at the interface between domains 1 and 2 of Cwp2 and Cwp8 that allowed domain 2 to assume an orientation closer to that seen in the other structure (Figure 6.9). In LMW SLP, the structure and orientation of domain 2 showed much greater differences than between Cwp2 and Cwp8, so it was not possible to observe the same orientation seen in the other two structures, however a hinging between the two domains was also seen.

Domain 1 of Cwp2 possesses a very similar fold to that of domain 1 of LMW SLP and Cwp8 (Figure 6.11) (Fagan *et al.*, 2009; Usenik *et al.*, 2017). In all three proteins, the domain is likely to be more buried than domain 2. Domain 1 of LMW SLP shows a higher degree of conservation across strains than domain 2. It is also therefore highly likely that they share a currently unidentified function.

Domain 3 of Cwp2 and Cwp8 (Figure 6.12) assumes a similar fold to domain 1, The structure of Cwp8 demonstrates that there is a significant amount of interaction between domain 3 and the first cell wall binding domain (CWBD). It therefore seems logical that the fold of this domain will be optimised for interaction with the CWBDs. It may, therefore, be that domain 3 of LMW SLP bears some similarity to that of Cwp2 and Cwp8, although, in the case of Cwp2 and Cwp8, the attachment to the cell wall binding domains is covalent, while the interaction of LMW SLP to HMW SLP is non-covalent (Fagan *et al.*, 2009).

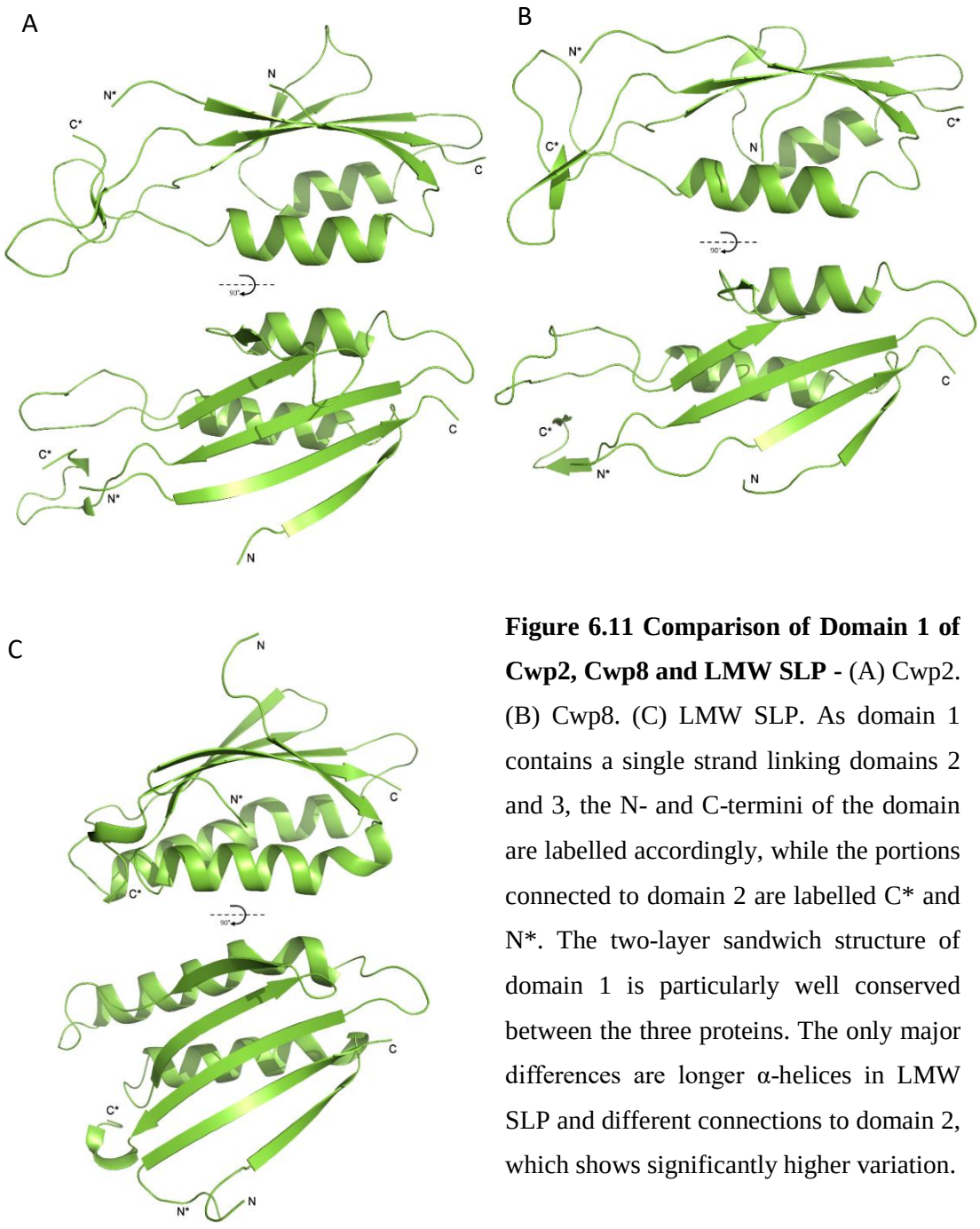


Figure 6.11 Comparison of Domain 1 of Cwp2, Cwp8 and LMW SLP - (A) Cwp2. (B) Cwp8. (C) LMW SLP. As domain 1 contains a single strand linking domains 2 and 3, the N- and C-termini of the domain are labelled accordingly, while the portions connected to domain 2 are labelled C* and N*. The two-layer sandwich structure of domain 1 is particularly well conserved between the three proteins. The only major differences are longer α-helices in LMW SLP and different connections to domain 2, which shows significantly higher variation.

Usenik *et al.* (2017) noted three portions of the functional region of Cwp8 that showed poor electron density in their structure, notably, two of these, A142-K151 and I189-G198, are found in domain 2 and are likely to form the most surface exposed areas of the protein *in vivo*. The equivalent regions in Cwp2, Lys141-Asp153 and Asn195-Lys218 are much more ordered. The latter of these two regions is also considerably longer in Cwp2, assumes a significantly different conformation and is stabilised by crystal packing while the former is ordered despite being solvent exposed. The third region identified in Cwp8 was an exposed loop in domain 3, S267-P273 (Figure 6.12). This loop also shows poor density in Cwp2 (Ile274-Tyr284), but is much shorter than in Cwp8. Interestingly, these three regions show the lowest degree of sequence similarity between the two proteins (figure 6.8).

While the attempt at phasing using models produced by rosetta was unsuccessful they do superpose well on the final structure of $Cwp2_{29\text{-}318}$ (figure 6.13). This shows that despite a lack of any similar structures, rosetta was able to produce two accurate models. It is therefore likely that molecular replacement with these models failed because they were too small, so Phaser was unable to place them correctly. It is possible, though that if the models had been clustered and ensembles produced from them, they may have just been close enough for molecular replacement to succeed. Despite the lack of success, this serves as an example of accurate *ab inito* modelling, which will surely become increasingly more common as fold space is filled and computers become more powerful.

As discussed in Chapter 2, data collection strategies have changed somewhat in recent years. While previously, data would have had to have been collected from a large number of crystals to obtain a single, complete dataset, the introduction of cryo-cooling 30 years ago resulted in a massive reduction of the effects of radiation damage, allowing a complete dataset to be collected from a single crystal (Hope, 1988). In more recent years, the introduction of better detectors and better programs for processing data has allowed strategies to be developed not to obtain a complete dataset, as this is now almost trivial, but to obtain a high multiplicity dataset, allowing much more precise determinations of intensities at low to medium resolution bins

and for information to be “squeezed out” from even the weakest reflections at high resolution.

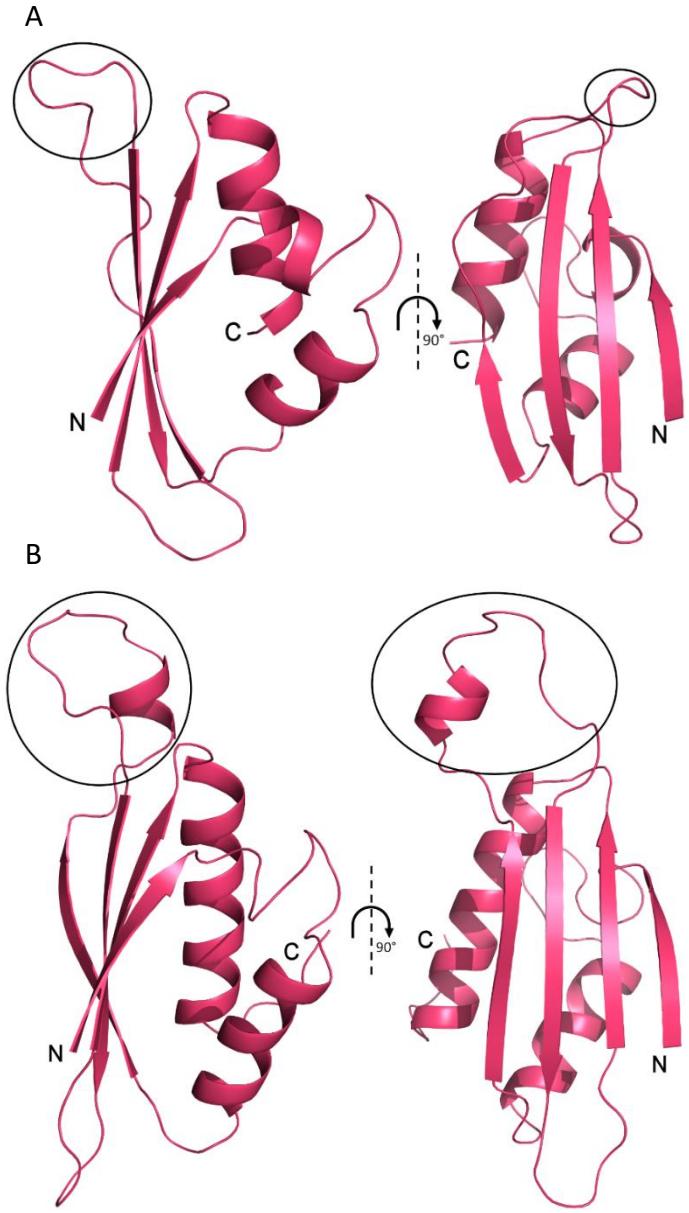


Figure 6.12 Comparison of Domain 3 of Cwp2, Cwp8 - (A) Cwp2. (B) Cwp8. In both proteins, domain 3 shows a similar two-layer sandwich fold to domain 1. The most noticeable difference between the two is the loop region protruding from the top of the domain (circled), this region is much longer in Cwp8 than in Cwp2. In both structures, this region had poor density.

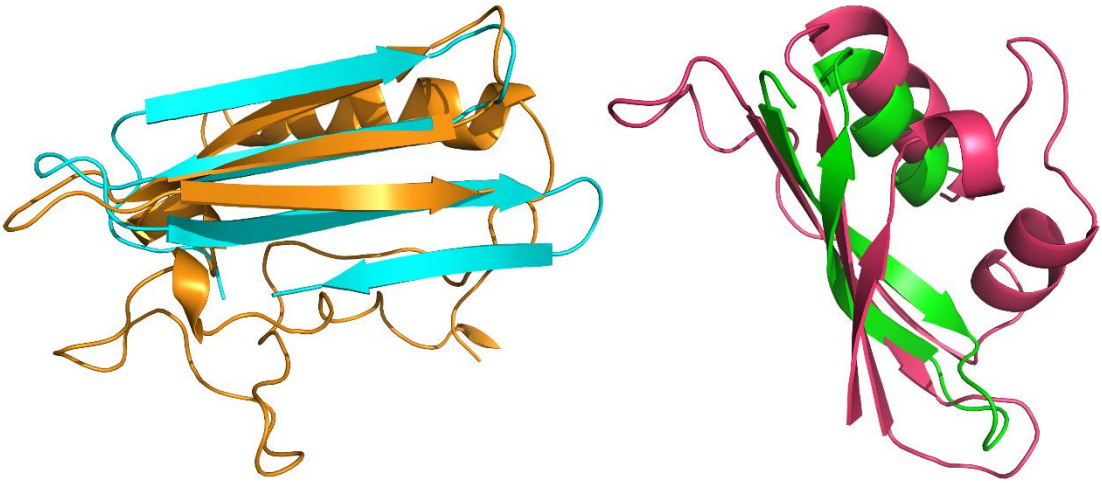


Figure 6.13 Domains 2 and 3 of Cwp2 with Rosetta models - The models superimpose on the structures fairly closely, but they were not sufficient for successful molecular replacement.

Higher multiplicity can result in significant improvements to CC1/2, $\langle |I|/\sigma(I) \rangle$ and the resolution of the data and will ultimately improve the resulting structure. Such a difference can be seen between the data presented in this thesis with $Cwp2_{27\text{-}322}$ and $Cwp19_{27\text{-}401}$ having higher multiplicities and Cwp84 datasets, which were collected earlier, having significantly lower multiplicities. This effect can also be observed by cutting a significant amount of data from a high multiplicity dataset (figure 6.14). Although it is still possible that the structure of $Cwp2_{27\text{-}322}$ could have been determined with lower multiplicity data, the high multiplicity has allowed the structure to be determined to a higher resolution with significant improvements to data at resolutions above approximately 3.3 Å. In many marginal cases, such as that of $Cwp2_{27\text{-}322}$, as indicated by the low translation function Z-scores, this may make all the difference between successful and unsuccessful structure determination. While this does not render older data obsolete, it does demonstrate the continuous improvement in technology, software and techniques that results in better data and more precise structures and potentially, a better chance of solving the structure in question.

It has been shown that LMW SLP and Cwp2 contribute somewhat towards the ability of *C. difficile* cells to adhere to host cells. Because of the high level of structural similarity between Cwp2 and Cwp8, this is also likely to be true for Cwp8. As domain

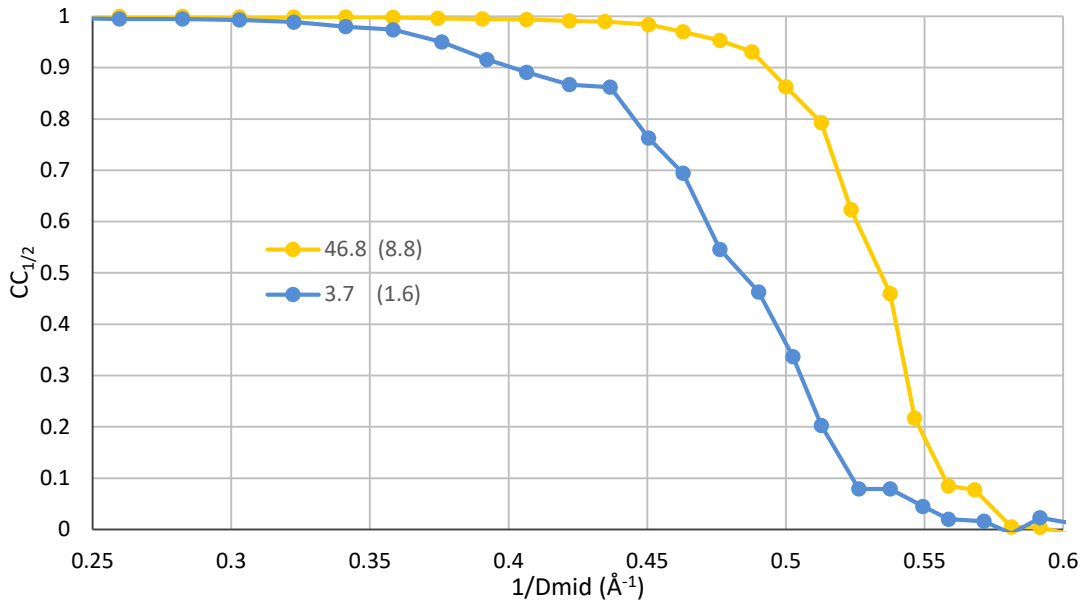


Figure 6.14 Comparison of high- and low-multiplicity data. The results from two instances of scaling of the Cwp2 data are presented with 1.65 Å cut-offs. One instance includes data from all 720° of images, while the other includes only 40° of data. Overall multiplicities are given with outer shell multiplicities in brackets. It can be seen that high multiplicity data produces a much smoother line and the data are better at all resolutions beyond approximately 1/Dmid = 0.3 (3.3 Å).

2 is the most surface exposed and variation in LMW SLP leads to changes in adhesive properties, it stands to reason that domain 2 is primarily responsible for adhesion. It has yet to be established though, how domain 2 performs this function. It is also unclear whether this role is solely carried out by domain 2 or whether it is assisted by domains 1 and 3.

The precise adhesive abilities of all three proteins need be more thoroughly explored using a range of mammalian cell lines. Do the different orientations of domain 2 affect adhesion? If domain 2 from Cwp2 were placed on Cwp8, (and *vice versa*) would the adhesive properties of the resulting protein be closer to those of former or the latter? Are the loop regions in domain 2 responsible for binding? Assuming it wouldn't affect the rest of the structure, would deleting these loop regions affect adhesion? If so, could this method be used to identify specific residues responsible for forming a binding site? Finally, could this information be used to disrupt binding?

Evidence now exists indicating roles in adhesion to host cells for LMW SLP (Merrigan *et al.*, 2013), Cwp66 (Waligora *et al.*, 2001), Cwp2, and, due to the similarity to Cwp2, Cwp8. Flagellar proteins FliC and FliD (Tasteyre *et al.*, 2001), GroEL (Hennequin *et al.*, 2001) fibronectin binding protein (Hennequin *et al.*, 2003) and lipoprotein CD0873 (Kovacs-Simon *et al.*, 2014), have also been linked to adhesion. This shows that adhesion to host cells is a multifactorial process and is unlikely to be completely disrupted by targeting a single protein. However, an in-depth understanding of the role of each of these proteins will be required to determine suitable targets. This work somewhat improves understanding of cellular adhesion by Cwp2 and, due to structural similarity, LMW SLP and Cwp8. It also possible that other proteins within the S-layer of *C. difficile* may possess roles in adhesion. Do they and Cwp66 bear structural similarities to these proteins, or do they use differing strategies?

Chapter 7

Discussion

7.1 Summary

Aside from serendipitous discoveries, such as the recent identification of a new bactericidal compound in soil samples by Maffioli *et al.* (2017), the successful development of novel drugs requires detailed structural knowledge of the intended targets alongside information from a range of other techniques (Chen *et al.*, 2012). If, therefore, drugs are to be developed to target the S-layer of *C. difficile*, the structures of the proteins therein are likely to be of major importance. The work presented in this thesis has resulted in the determination of near-atomic resolution crystal structures of the functional domains of Cwp84 with its propeptide and Cwp19, while high resolution structures have also been determined for the functional domains of Cwp84 without its propeptide and Cwp2. Based on these structures and the recently determined structures of Cwp6 and Cwp8 (Usenik *et al.*, 2017), figure 1.4 can be updated to figure 7.1.

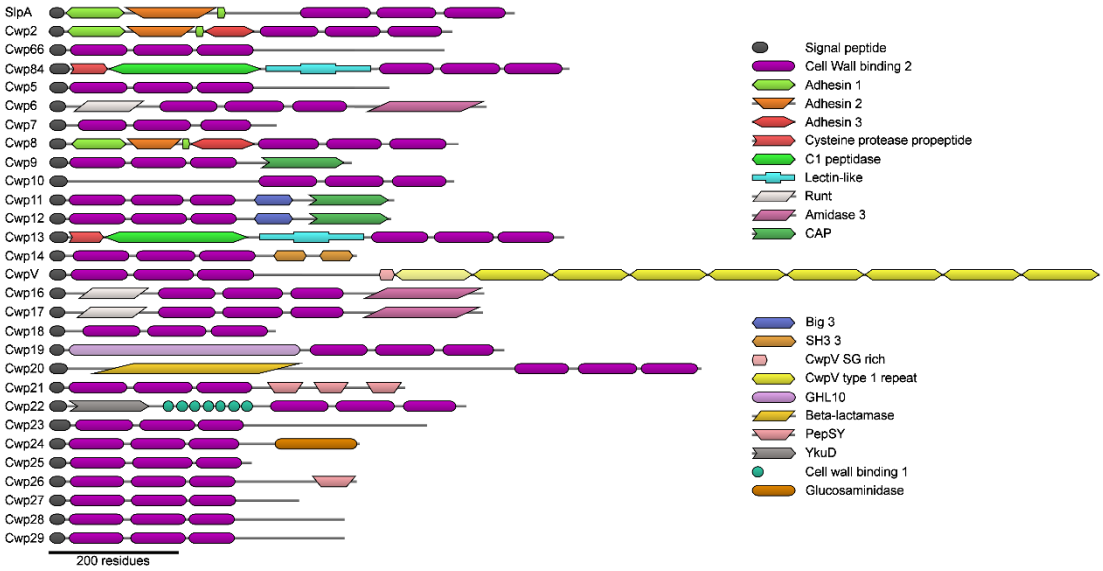


Figure 7.1 Domain representation of the 29 cwp genes found in the Clostridium difficile 630 genome. The figure is an updated version of the representation given in figure 1.4. Changes have been made to SlpA, Cwp2, Cwp84, Cwp6, Cwp8, Cwp13, Cwp16, Cwp17 and Cwp19. A significant amount of information relating to the structure of the S-layer has been determined in recent years, however, there is still a much larger number of structures yet to be determined.

7.1.1 Cwp84

Cwp84 was determined to possess a Cathepsin L-like cysteine protease domain, a newly identified lectin-like domain and a novel propeptide fold. After cleavage of the propeptide, conformational changes were observed within the active site groove of Cwp84 and on the surface of the lectin-like domain. It is likely that the conformational changes within the active site groove facilitate binding of SlpA, but structures with SlpA and inhibitors bound will be required to fully understand this process.

The Lectin like domain forms part of the S end of the cysteine protease active site. BLAST searches revealed that lectin-like domains are paired with a cysteine protease domain in a large number of bacterial proteins, indicating that the presence of the former is crucial to the functioning of the latter. Some of these proteins possess cell wall binding domains while others do not. The precise role of the lectin-like domain is yet to be determined, however, as is any actual ability to bind carbohydrates. If the lectin-like domain does indeed bind carbohydrates, determination of the precise ligands it is capable of binding and the effect of binding on catalysis may yield information on the overall function of the domain. The positioning of the lectin-like domain in any future structures of complexes of Cwp84 and SlpA may also shed light on the role of the lectin-like domain.

Cwp84 and Cwp13 share 63% identity and 78% similarity (strain 630) (Altschul *et al.*, 1990). Despite this, it has been demonstrated that Cwp13 is unable to cleave SlpA in the same way as Cwp84 and may possess a role in “cleaning up” misfolded protein (de la Riva *et al.*, 2011). It goes without saying the seemingly minor structural differences between the two proteins will result in the observed major functional differences. Future structural characterisation of Cwp13 and of Cwp84 with inhibitors bound, potentially ones that mimic the substrate, should therefore reveal the exact nature of these crucial structural differences, which will be of major importance for the design of inhibitors specific to Cwp84.

7.1.2 Cwp19

Cwp19, which belongs to the recently identified glycoside hydrolase-like family 10 (GHL10), was shown to possess a TIM barrel fold, which is common among many

glycoside hydrolases (GH). The location of the active site was determined and probable active site residues were identified. GH activity was demonstrated by the ability to break down peptidoglycan resulting in the lysis of *Micrococcus luteus* cells. The peptidoglycan hydrolase activity was determined to be an order of magnitude slower than that of lysozyme. Work to determine the precise glycosidic bonds that Cwp19 is capable of breaking down was unsuccessful, but may demonstrate a high degree of substrate selectivity.

If Cwp19 is involved in the formation of PSII, as the positioning of the gene within the AP locus suggests, determination of the precise function of Cwp19, and indeed of the other proteins coded for by the locus, will be of major importance. This will require knockouts, more activity assays and structures with substrates bound. These can be accompanied by kinetics studies and mutations to determine precise mechanisms and residues involved in catalysis. These techniques, combined with structural studies, could be used to design inhibitors to disrupt PSII synthesis.

7.1.3 Cwp2

The functional portion of Cwp2 was shown to possess a three domain fold with a high level of similarity to that of Cwp8 and significant similarity to LMW SLP. Work performed by Jon Kirby demonstrated adhesive properties of Cwp2 *in vitro*, which has also been demonstrated for LMW SLP and Cwp66 (Waligora *et al.*, 2001; Merrigan *et al.*, 2013). The genes coding for these four proteins have been demonstrated to exhibit the most variation within the SlpA locus (Dingle *et al.*, 2013). Domain 2 of LMW SLP, Cwp2, and Cwp8 is likely to be the most surface exposed and therefore primarily responsible for the adhesive properties. This domain also shows significant variation between the three proteins, which could be linked to variability in binding modes to host cells or immune system evasion.

Further work needs to be done to determine how these proteins bind to host cells. This information could eventually be used to disrupt binding of *C. difficile* to host cells. Binding to host cells, is however, a multifactorial process. While disruption of these proteins alone is unlikely to completely prevent binding, it may still have a significant

effect. It may also be possible to combine disruption of binding mediated by the S-layer with other targets.

7.2 Cell Wall Binding Domains

The three proteins studied in this thesis, along with up to 26 others from *C. difficile* each possess three cell wall binding domains. These domains mediate binding to the cell wall through a non-covalent interaction with the surface bound polysaccharide PSII (Ganeshapillai *et al.*, 2008; Willing *et al.*, 2015; Chu *et al.*, 2016). It follows logically from this that targeting the binding of PSII would be a sensible way of attacking *C. difficile*. This will require structures of the cell wall binding domains and an understanding of their interactions with PSII, work to this end was performed but it was largely unsuccessful.

Full length structures were recently published for Cwp6 and Cwp8 (Usenik *et al.*, 2017). These structures revealed that the three cell wall binding domains each possess a toprim fold (named after topoisomerase primase) formed by a four-stranded β-sheet sandwiched between two layers each containing two α-helices (figure 7.2A). Three α-helices, one from each domain, come together to form a central axis. Each of these domains is rotated 120° relative to the other two around the central axis in a trefoil-like arrangement. The previously identified PILL motif, which was thought to be involved in binding PSII (Willing *et al.*, 2015), has been determined instead to be crucial to the formation of an interface binding the inner most β-strand of one domain to the preceding domain and to the central axis.

Importantly, these structures demonstrated that the portions of the sequence that were thought to correspond to each domain were somewhat incorrect (figure 7.2A). Although the second and third cell wall binding domains are essentially consistent with the predicted domains, there are significant differences in the first domain. Taking Cwp8 as an example, which possesses a very similar fold to that of Cwp2 (figure 6.9), the cell wall binding domains were predicted to correspond to residues 323-417, 428-513 and 525-613. It is now evident that residues 291-310 form the final β-strand and α-helix of what could be considered the third cell wall binding domain,

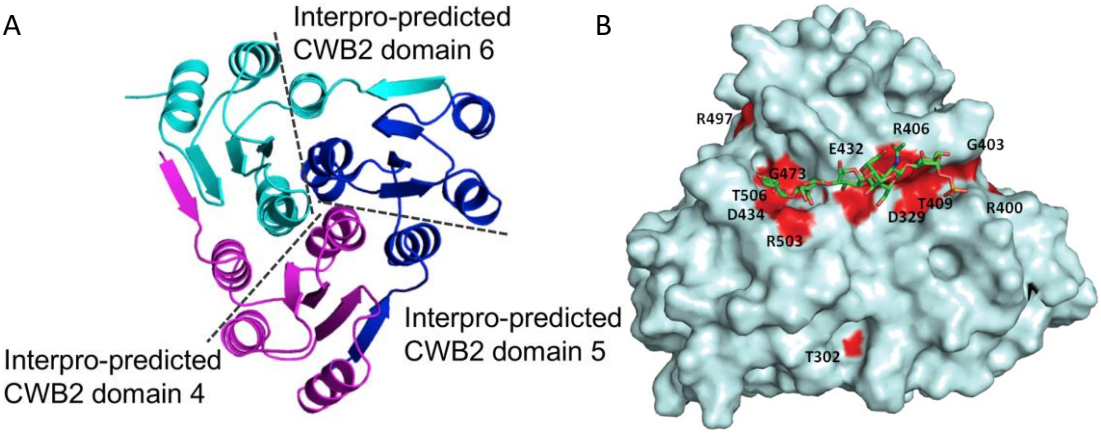


Figure 7.2 Cell wall binding domain structures. The figures are taken from (Usenik *et al.*, 2017). (A) Comparison between the predicted and identified domains of Cwp8. The predicted domains are coloured separately while the identified domains are separated by dashed lines. (B) Docking of PSII to the CWB2 trefoil. One of the two suggested binding modes is shown. Residues conserved across all 29 Cwps are highlighted in red. The predicted binding interface is formed by all three CWB2 domains.

residues 318-416 form the first domain, residues 423-515 form the second domain and residues 521-596 form the majority of the third domain. This is the reason why constructs 2 and 3 in Chapter 4 were flawed - they will have contained part of the third cell wall binding domain, which will have, at best, caused problems with correct folding and, at worst, completely prevented it. Constructs 5 and 6 likely will have contained the correct CWB2 domains, but, as discussed in the following paragraph, all three may be required.

A domain is defined as "A discrete, independent folding unit within the tertiary structure of a protein" (Horton *et al.*, 2006). If the cell wall binding "domains" are taken as discrete, the hydrophobic PILL motif would be found on the surface of each of the three domains and only buried upon trimerisation. Moreover, CWB2 domains always appear to be found in threes. This, and the fact that constructs in chapter 4 with one or two cell wall binding domains did not crystallise may suggest that all three domains are required for correct folding. Should the cell wall binding domains, therefore, be considered three separate domains that trimerise, or should they be considered one single domain with three repeat regions? This distinction is especially evident when considering that domain 3, whose primary sequence is separated by

that of domains 1 and 2, would be unlikely to fold correctly without the latter two present. On top of this, the actual interface to which Usenik *et al.* (2017) modelled the binding of PSII is formed on the face of the trefoil by residues from all three domains, demonstrating that they would be unable to function individually (figure 7.2B).

7.3 Future Work

For reasons given in the previous section, any future constructs should have either all or none of the cell wall binding repeats, which should potentially be considered a single domain. The location of these domains/repeats within the protein and the effect this has on the positioning of any functional domains should also be considered. Fourteen Cwps possess N-terminal CWB2 domains, nine have C-terminal domains, in three they are central and three proteins appear to consist of little but the three CWB2 domains. How does this differing positioning of the CWB2 domains affect the positioning of the functional domains of the protein? Cwp6, Cwp16 and Cwp17 have central CWB2 domains with a C-terminal amidase domain. Considering the size of the protein and that it will be surrounded by the more extended structure of LMW SLP, it seems unlikely that they would be exposed enough at the surface to have an effect on other cells. Does this mean that the amidase domains are positioned in such a way that they are able to process the peptidoglycan of *C. difficile*? If this is the case, where would the runt domains be positioned? Biazzo *et al.* (2013) demonstrated that due to conservation of sequence and expression patterns Cwp16 is likely to possess an important function but Cwp17 isn't. What differences exist between these two proteins that result in this discrepancy and what might this important role be?

If C-terminal CWB2 domains result in the surface exposure of the functional region, how does the presence of N-terminal cell wall binding domains affect the functional regions? While the N- and C-termini of the CWB2 trefoil are relatively close, their separation could affect the positioning of the functional regions. Cwp66, the only Cwp with N-terminal CWB2 domains to be studied to any significant degree, has been

shown to possess adhesive properties (Waligora *et al.*, 2001). This must mean that the functional region is surface exposed. Does this mean that the three domains differ depending on their positioning within the primary sequence, which would allow somewhat different binding to PSII and therefore, different orientation or does Cwp66 possess an extended loop that allows the functional region to be surface exposed? Usenik *et al.* (2017) considered the binding of PSII by the cell wall binding domains but did not thoroughly analyse the differences in binding between Cwp6 and Cwp8. It has been shown that different methods of S-layer extraction yield different amounts of certain proteins (Wright *et al.*, 2005), is this a result of differences in PSII binding? As PSII binding is essential for the formation of the S-layer, this would seem like a sensible thing to attempt to disrupt. Ideally, multiple Cwps could be targeted simultaneously to prevent resistance developing. The exact method by which each protein being targeted binds to PSII should be thoroughly analysed. If there are significant differences, any inhibitor would need to either bind to multiple different proteins, or would need to have a significant number of variants to bind to different proteins. The issues encountered in this thesis should no longer be a problem, as a method has been demonstrated for the soluble expression of full length Cwps (Usenik *et al.*, 2017).

The structure of $Cwp84_{33\text{-}497\_C116A}$ was determined by Selenium-MAD, $Cwp19_{27\text{-}401}$ used Se-SAD and $Cwp2_{27\text{-}322}$ required non-trivial molecular replacement using models derived from Cwp8. The determination of the structures of other proteins from the S-layer of *C. difficile* has also required more complex phasing methods. LMW SLP was determined using Se-MAD (Fagan *et al.*, 2009), Cwp8 used a platinum derivative and the structure of Cwp6 was determined by molecular replacement using models based on the cell wall binding domains of Cwp8 and an amidase from *Bacillus polymyxa* (Usenik *et al.*, 2017). This demonstrates not only how proteins within the S-layer differ significantly from other structures, but also how they differ significantly from each other. This is discussed in Chapter 6 with respect to LMW SLP, Cwp2 and Cwp8 and earlier in this chapter with respect to Cwp84 and Cwp13, but the differences between other proteins should also be considered. If this trend of "structural isolation" is shown by other proteins with the S-layer, the future determination of

other structures could very well yield unexpected results and leads to many potential questions:

Is the degree of difference between LMW SLP, Cwp2 and Cwp8 reflected by Cwp6, Cwp16 and Cwp17? Do their functions differ based upon these differences? How do the Big 3 and CAP domains of Cwp11 and Cwp12 differ from each other? Does any variation between them affect their functions and how is the CAP domain of Cwp9 affected by the lack of a Big 3 domain? Similar questions can be asked for the differing number of PepSY domains in Cwp21 and Cwp26.

Two proteins that warrant particular note are Cwp66 and CwpV. Cwp66 has been shown to have a role in adhesion to host cells but the mechanism by which this occurs is completely unknown (Waligora *et al.*, 2001). Does it bear any structural similarity to LMW SLP, Cwp2 and Cwp8, or does it adhere using an entirely different fold? The presence of three imperfect repeat regions, which are not seen in the other three proteins would suggest the latter. Determining the structure of Cwp66 would be a good starting point for further characterising these adhesive properties, which could then be further analysed by methods similar to those suggested for Cwp2 and Cwp8.

CwpV is a particularly unusual protein. While the sequence is well conserved up to the SG rich region, the repeat regions after this have been shown to possess five different completely unrelated possible sequences that all carry out the same function (Reynolds *et al.*, 2011). How this variability came about is unclear, but more importantly, so too is how all five repeat types are able to confer the same function. CwpV mediates aggregation of *C. difficile* cells and is expressed in a relatively small proportion of cells. Structures of the five different repeat regions may well answer questions about how the protein is able to function and could potentially be used to disrupt aggregation and prevent *C. difficile* from colonising the gut.

Many of the Cwps are yet to be thoroughly characterised. Although many different domains have been predicted to be present and potential roles for each protein can be suggested, as in Chapter 1, a study of each protein is required to fully understand their functions, how essential each protein is and therefore, how suitable each one

would be as a drug target. Knockouts have been performed on some (Kirby *et al.*, 2009; Kirby, 2011; Dembek, 2014; Bradshaw *et al.*, 2017a) but even many of these lack thorough analysis. As discussed in Chapter 6, Cwp2 and Cwp8 knockouts could be performed along with mutations to the two proteins, the formation of chimeras and more thorough adhesion assays to more precisely determine the role of each of the two proteins and their domains in binding to the host. Similar thorough knockout studies should be performed on each of the other Cwps to improve understanding of their roles. These studies can be accompanied by structural, kinetic and mutation studies to confirm and further characterise any putative functions. It is also likely that this would lead to many unpredictable results and new directions that research could progress in. As mentioned in Chapter 1, eight Cwps "each contain regions of around 100 residues or more for which no structure or function has so far been predicted. This leaves a large number of potential functions of the S-layer still to be determined." Structural characterisation in combination with knockout work seems like a good starting point for the determination of roles for each of these proteins, which would likely result in the identification of new functions for the S-layer of *C. difficile* and many potential new drug targets for the prevention and treatment of CDI.

References

Abraham EP, Chain E (1940) "An Enzyme from Bacteria able to Destroy Penicillin", Nature **146**: 873.

Abrahams JP, Leslie AG (1996) "Methods used in the structure determination of bovine mitochondrial F1 ATPase", Acta Cryst. D **52**(Pt 1): 30-42.

Adams-Cioaba MA, Krupa JC, Xu C, Mort JS, Min J (2011) "Structural basis for the recognition and cleavage of histone H3 by cathepsin L", Nat. Comms. **2**: 197.

Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Zwart PH (2010) "PHENIX: a comprehensive Python-based system for macromolecular structure solution", Acta Cryst. D **66**(Pt 2): 213-221.

Alber T, Banner DW, Bloomer AC, Petsko GA, Phillips D, Rivers PS, Wilson IA (1981) "On the three-dimensional structure and catalytic mechanism of triose phosphate isomerase", Philos. Trans. R. Soc. Lond., Ser. B: Biol. Sci. **293**(1063): 159-171.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) "Basic local alignment search tool", J. Mol. Biol. **215**(3): 403-410.

Arndt UW, Crowther RA, Mallett JF (1968) "A computer-linked cathode-ray tube microdensitometer for x-ray crystallography", J. Sci. Inst. **1**(5): 510-516.

Asensio A, Bouza E, Grau S, Rubio-Rodriguez D, Rubio-Terres C (2013) "[Cost of *Clostridium difficile* associated diarrhea in Spain]", Rev. Esp. Sal. Pub. **87**(1): 25-33.

Assmann G, Brehm W, Diederichs K (2016) "Identification of rogue datasets in serial crystallography", J. Appl. Cryst. **49**(Pt 3): 1021-1028.

Baerends RJ, Smits WK, de Jong A, Hamoen LW, Kok J, Kuipers OP (2004) "Genome2D: a visualization tool for the rapid analysis of bacterial transcriptome data", Genome Biol. **5**(5): R37.

Barra-Carrasco J, Paredes-Sabja D (2014) "*Clostridium difficile* spores: a major threat to the hospital environment", Fut. Microbiol. **9**(4): 475-486.

Barrett AJ, Kembhavi AA, Brown MA, Kirschke H, Knight CG, Tamai M, Hanada K (1982) "L-trans-Epoxysuccinyl-leucylamido(4-guanidino)butane (E-64) and its analogues as inhibitors of cysteine proteinases including cathepsins B, H and L", Biochem. J. **201**(1): 189-198.

Bateman A, Eddy SR, Chothia C (1996) "Members of the immunoglobulin superfamily in bacteria", Prot. Sci. **5**(9): 1939-1941.

Battye TG, Kontogiannis L, Johnson O, Powell HR, Leslie AG (2011) "iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM", Acta Cryst. D **67**(Pt 4): 271-281.

Beevers CA, Lipson H (1934) "The Crystal Structure of Copper Sulphate Pentahydrate, $CuSO_4.5H_2O$", Proc. R. Soc. A: Mat. Phys. Eng. Sci. **146**(858): 570-582.

Benedict SR (1909) "A reagent for the detection of reducing sugars", J. Biol. Chem. **5**(5): 485-487.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) "The Protein Data Bank", Nucleic Acids Res. **28**(1): 235-242.

Bertani G (1951) "Studies on lysogenesis. I. The mode of phage liberation by lysogenic *Escherichia coli*", J. Bacteriol. **62**(3): 293-300.

Beton D, Guzzo CR, Ribeiro AF, Farah CS, Terra WR (2012) "The 3D structure and function of digestive cathepsin L-like proteinases of *Tenebrio molitor* larval midgut", Insect Biochem. Mol. Biol. **42**(9): 655-664.

Biarrotte-Sorin S, Hugonnet JE, Delfosse V, Mainardi JL, Gutmann L, Arthur M, Mayer C (2006) "Crystal structure of a novel beta-lactam-insensitive peptidoglycan transpeptidase", J. Mol. Biol. **359**(3): 533-538.

Biazzo M, Cioncada R, Fiaschi L, Tedde V, Spigaglia P, Mastrantonio P, Pizza M, Barocchi MA, Scarselli M, Galeotti CL (2013) "Diversity of cwp loci in clinical isolates of *Clostridium difficile*", J. Med. Microbiol. **62**(Pt 9): 1444-1452.

Bielnicki J, Devedjiev Y, Derewenda U, Dauter Z, Joachimiak A, Derewenda ZS (2006) "*B. subtilis* ykuD protein at 2.0 A resolution: insights into the structure and function of a novel, ubiquitous family of bacterial enzymes", Proteins **62**(1): 144-151.

Blakeley MP, Hasnain SS, Antonyuk SV (2015) "Sub-atomic resolution X-ray crystallography and neutron crystallography: promise, challenges and potential", IUCrJ **2**(Pt 4): 464-474.

Blow DM (2002) "Outline of crystallography for biologists", Oxford University Press, Oxford ; New York

Bond AD (2015) "Why do we trust X-ray crystallography?", Resonance **19**(12): 1087-1092.

Bond CS, Schuttelkopf AW (2009) "ALINE: a WYSIWYG protein-sequence alignment editor for publication-quality alignments", Acta Cryst. D **65**(Pt 5): 510-512.

Bonneau R, Strauss CE, Rohl CA, Chivian D, Bradley P, Malmstrom L, Robertson T, Baker D (2002) "De novo prediction of three-dimensional structures for major protein families", J. Mol. Biol. **322**(1): 65-78.

Borchardt-Ott W (2011) "Crystallography: An Introduction", Springer, Germany

Borgia G, Maraolo AE, Foggia M, Buonomo AR, Gentile I (2015) "Fecal microbiota transplantation for *Clostridium difficile* infection: back to the future", Expert Opin. Biol. Ther. **15**(7): 1001-1014.

Bourgault R, Oakley AJ, Bewley JD, Wilce MC (2005) "Three-dimensional structure of (1,4)-beta-D-mannan mannanohydrolase from tomato fruit", Prot. Sci. **14**(5): 1233-1241.

Boyd WC (1954) The Proteins of Immune Reactions. The Proteins. K Bailey, Academic Press**:** 755-844.

Bradshaw WJ, Kirby JM, Thiyagarajan N, Chambers CJ, Davies AH, Roberts AK, Shone CC, Acharya KR (2014) "The structure of the cysteine protease and lectin-like domains of Cwp84, a surface layer-associated protein from *Clostridium difficile*", Acta Cryst. D **70**(Pt 7): 1983-1993.

Bradshaw WJ, Roberts AK, Shone CC, Acharya KR (2015) "Cwp84, a *Clostridium difficile* cysteine protease, exhibits conformational flexibility in the absence of its propeptide", Acta Cryst. F **71**(Pt 3): 295-303.

Bradshaw WJ, Kirby JM, Roberts AK, Shone CC, Acharya KR (2017a) "Cwp2 from *Clostridium difficile* exhibits an extended three domain fold and cell adhesion in vitro", FEBS J. **284**(17): 2886-2898.

Bradshaw WJ, Rehman S, Pham TT, Thiyagarajan N, Lee RL, Subramanian V, Acharya KR (2017b) "Structural insights into human angiogenin variants implicated in Parkinson's disease and Amyotrophic Lateral Sclerosis", Sci. Rep. **7**: 41996.

Bragg WL (1913) "The Structure of Some Crystals as Indicated by Their Diffraction of X-rays", Proc. R. Soc. A: Mat. Phys. Eng. Sci. **89**(610): 248-277.

Brangulis K, Jaudzems K, Petrovskis I, Akopjana I, Kazaks A, Tars K (2015) "Structural and functional analysis of BB0689 from *Borrelia burgdorferi*, a member of the bacterial CAP superfamily", J. Struct. Biol **192**(3): 320-330.

Brunger AT (1992) "Free R value: a novel statistical quantity for assessing the accuracy of crystal structures", Nature **355**(6359): 472-475.

Calabi E, Ward S, Wren B, Paxton T, Panico M, Morris H, Dell A, Dougan G, Fairweather N (2001) "Molecular characterization of the surface layer proteins from *Clostridium difficile*", Mol. Microbiol. **40**(5): 1187-1199.

Calabi E, Fairweather N (2002) "Patterns of sequence conservation in the S-Layer proteins and related sequences in *Clostridium difficile*", J. Bacteriol. **184**(14): 3886-3897.

Carrion AF, Hosein PJ, Cooper EM, Lopes G, Pelaez L, Rocha-Lima CM (2010) "Severe colitis associated with docetaxel use: A report of four cases", World J. Gast. Onc. **2**(10): 390-394.

CCP4 (1994) "The CCP4 suite: programs for protein crystallography", Acta Cryst. D **50**(Pt 5): 760-763.

Cerquetti M, Pantosti A, Stefanelli P, Mastrantonio P (1992) "Purification and characterization of an immunodominant 36 kDa antigen present on the cell surface of *Clostridium difficile*", Microbiol. Path. **13**(4): 271-279.

Cerquetti M, Molinari A, Sebastianelli A, Diociaiuti M, Petruzzelli R, Capo C, Mastrantonio P (2000) "Characterization of surface layer proteins from different *Clostridium difficile* clinical isolates", Microbiol. Path. **28**(6): 363-372.

ChapetónMontes D, Candela T, Collignon A, Janoir C (2011) "Localization of the *Clostridium difficile* cysteine protease Cwp84 and insights into its maturation process", J. Bacteriol. **193**(19): 5314-5321.

Chen L, Morrow JK, Tran HT, Phatak SS, Du-Cuny L, Zhang S (2012) "From laptop to benchtop to bedside: structure-based drug design on protein targets", Curr. Pharm. Des. **18**(9): 1217-1239.

Chen VB, Arendall WB, 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010) "MolProbity: all-atom structure validation for macromolecular crystallography", Acta Cryst. D **66**(Pt 1): 12-21.

Cheng Y (2015) "Single-Particle Cryo-EM at Crystallographic Resolution", Cell **161**(3): 450-457.

Chothia C, Lesk AM (1986) The use of sequence homologies to predict protein structure. Computer Graphics and Molecular Modelling. R Fletterick, M Zoller. USA, Cold Spring Harbor Laboratory**:** 33-38.

Chu M, Mallozzi MJ, Roxas BP, Bertolo L, Monteiro MA, Agellon A, Viswanathan VK, Vedantam G (2016) "A *Clostridium difficile* Cell Wall Glycopolymer Locus Influences Bacterial Shape, Polysaccharide Production and Virulence", PLoS Path. **12**(10): e1005946.

Chumbler NM, Rutherford SA, Zhang ZF, Farrow MA, Lisher JP, Farquhar E, Giedroc DP, Spiller BW, Melnyk RA, Lacy DB (2016) "Crystal structure of *Clostridium difficile* toxin A", Nat. Microbiol. **1**(1): 15002.

Coelho LC, Silva PM, Lima VL, Pontual EV, Paiva PM, Napoleao TH, Correia MT (2017) "Lectins, Interconnecting Proteins with Biotechnological/Pharmacological and Therapeutic Applications", Evid. Based Compl. Alt. Med. **2017**: 1594074.

Costello SP, Conlon MA, Vuaran MS, Roberts-Thomson IC, Andrews JM (2015) "Faecal microbiota transplant for recurrent *Clostridium difficile* infection using long-term frozen stool is effective: clinical efficacy and bacterial viability data", Aliment. Pharmacol. Ther. **42**(8): 1011-1018.

Coulombe R, Grochulski P, Sivaraman J, Menard R, Mort JS, Cygler M (1996) "Structure of human procathepsin L reveals the molecular basis of inhibition by the prosegment", EMBO J. **15**(20): 5492-5503.

Cowtan K (2006) "The *Buccaneer* software for automated model building. 1. Tracing protein chains", Acta Cryst. D **62**(Pt 9): 1002-1011.

Cowtan K (2010) "Recent developments in classical density modification", Acta Cryst. D **66**(Pt 4): 470-478.

Cowtan KD, Zhang KY (1999) "Density modification for macromolecular phase improvement", Prog. Biophys. Mol. Biol. **72**(3): 245-270.

Czibener C, Ugalde JE (2012) "Identification of a unique gene cluster of *Brucella* spp. that mediates adhesion to host cells", Microb. Infect. **14**(1): 79-85.

Dahl SW, Halkier T, Lauritzen C, Dolenc I, Pedersen J, Turk V, Turk B (2001) "Human Recombinant Pro-dipeptidyl Peptidase I (Cathepsin C) Can Be Activated by Cathepsins L and S but Not by Autocatalytic Processing", Biochem. **40**(6): 1671-1678.

Dang TH, de la Riva L, Fagan RP, Storck EM, Heal WP, Janoir C, Fairweather NF, Tate EW (2010) "Chemical probes of surface layer biogenesis in *Clostridium difficile*", ACS Chem. Biol. **5**(3): 279-285.

Dauter Z, Dauter M, Dodson E (2002) "Jolly SAD", Acta Cryst. D **58**(Pt 3): 494-506.

Davies AH, Roberts AK, Shone CC, Acharya KR (2011) "Super toxins from a super bug: structure and function of *Clostridium difficile* toxins", Biochem. J. **436**(3): 517-526.

de Graaff RA, Hilge M, van der Plas JL, Abrahams JP (2001) "Matrix methods for solving protein substructures of chlorine and sulfur from anomalous data", Acta Cryst. D **57**(Pt 12): 1857-1862.

de la Riva L, Willing SE, Tate EW, Fairweather NF (2011) "Roles of cysteine proteases Cwp84 and Cwp13 in biogenesis of the cell wall of *Clostridium difficile*", J. Bacteriol. **193**(13): 3276-3285.

Dembek M, Reynolds CB, Fairweather NF (2012) "*Clostridium difficile* cell wall protein CwpV undergoes enzyme-independent intramolecular autoproteolysis", J. Biol. Chem. **287**(2): 1538-1544.

Dembek M (2014) "Whole-genome analysis of sporulation and germination in *Clostridium difficile*", **PhD**, Imperial College London, London, UK.

Derewenda ZS, Vekilov PG (2006) "Entropy and surface engineering in protein crystallization", Acta Cryst. D **62**(Pt 1): 116-124.

Desvaux M, Dumas E, Chafsey I, Hebraud M (2006) "Protein cell surface display in Gram-positive bacteria: from single protein to macromolecular protein structure", FEMS Microbiol. Ecol. **256**(1): 1-15.

Diederichs K, Karplus PA (1997) "Improved R-factors for diffraction data analysis in macromolecular crystallography", Nat. Struct. Biol. **4**(4): 269-275.

Diederichs K, Karplus PA (2013) "Better models by discarding data?", Acta Cryst. D **69**(Pt 7): 1215-1222.

Dingle KE, Didelot X, Ansari MA, Eyre DW, Vaughan A, Griffiths D, Ip CL, Batty EM, Golubchik T, Bowden R, Jolley KA, Hood DW, Fawley WN, Walker AS, Peto TE, Wilcox MH, Crook DW (2013) "Recombinational switching of the *Clostridium difficile* S-layer and a novel glycosylation gene cluster revealed by large-scale whole-genome sequencing", J. Infect. Dis. **207**(4): 675-686.

Dodson E, Kleywegt GJ, Wilson K (1996) "Report of a workshop on the use of statistical validators in protein X-ray crystallography", Acta Cryst. D **52**(Pt 1): 228-234.

Dodson E (2003) "Is it jolly SAD?", Acta Cryst. D **59**(Pt 11): 1958-1965.

Drenth J, Jansonius JN, Koekoek R, Swen HM, Wolthers BG (1968) "Structure of papain", Nature **218**(5145): 929-932.

Drickamer K (1995) "Increasing diversity of animal lectin structures", Curr. Opin. Struct. Biol. **5**(5): 612-616.

Drickamer K (1999) "C-type lectin-like domains", Curr. Opin. Struct. Biol. **9**(5): 585-590.

Driessen AJ, Nouwen N (2008) "Protein translocation across the bacterial cytoplasmic membrane", Annu. Rev. Biochem. **77**: 643-667.

du Plessis DJ, Nouwen N, Driessen AJ (2011) "The Sec translocase", Biochim. Biophys. Acta **1808**(3): 851-865.

Durst KL, Hiebert SW (2004) "Role of RUNX family members in transcriptional repression and gene silencing", Oncogene **23**(24): 4220-4224.

Eddy SR (2008) "A probabilistic model of local sequence alignment that simplifies statistical significance estimation", PLoS Comp. Biol. **4**(5): e1000069.

Emerson J, Fairweather N (2009) Surface structures of *C. difficile* and other clostridia: Implications for pathogenesis and immunity. Clostridia: Molecular biology in the post-genomic era. H Bruggemann, G Gottschalk, Caister Academic Press**:** 157-167.

Emerson JE, Stabler RA, Wren BW, Fairweather NF (2008) "Microarray analysis of the transcriptional responses of *Clostridium difficile* to environmental and antibiotic stress", J. Med. Microbiol. **57**(Pt 6): 757-764.

Emerson JE, Reynolds CB, Fagan RP, Shaw HA, Goulding D, Fairweather NF (2009) "A novel genetic switch controls phase variable expression of CwpV, a *Clostridium difficile* cell wall protein", Mol. Microbiol. **74**(3): 541-556.

Emsley P, Cowtan K (2004) "Coot: model-building tools for molecular graphics", Acta Cryst. D **60**(Pt 12): 2126-2132.

Engelhardt H, Peters J (1998) "Structural research on surface layers: a focus on stability, surface layer homology domains, and surface layer-cell wall interactions", J. Struct. Biol **124**(2-3): 276-302.

Engh RA, Huber R (1991) "Accurate Bond and Angle Parameters for X-Ray Protein-Structure Refinement", Acta Cryst. A **47**(4): 392-400.

Evans G, Pettifer RF (2001) "CHOOCH: a program for deriving anomalous-scattering factors from X-ray fluorescence spectra", J. Appl. Cryst. **34**(1): 82-86.

Evans P (2006) "Scaling and assessment of data quality", Acta Cryst. D **62**(Pt 1): 72-82.

Evans PR, Murshudov GN (2013) "How good are my data and what is the resolution?", Acta Cryst. D **69**(Pt 7): 1204-1214.

Fagan RP, Albesa-Jove D, Qazi O, Svergun DI, Brown KA, Fairweather NF (2009) "Structural insights into the molecular organization of the S-layer from *Clostridium difficile*", Mol. Microbiol. **71**(5): 1308-1322.

Fagan RP, Fairweather NF (2011) "*Clostridium difficile* has two parallel and essential Sec secretion systems", J. Biol. Chem. **286**(31): 27483-27493.

Fagan RP, Janoir C, Collignon A, Mastrantonio P, Poxton IR, Fairweather NF (2011) "A proposed nomenclature for cell wall proteins of *Clostridium difficile*", J. Med. Microbiol. **60**(Pt 8): 1225-1228.

Fagan RP, Fairweather NF (2014) "Biogenesis and functions of bacterial S-layers", Nat. Rev. Microbiol. **12**(3): 211-222.

Feltcher ME, Braunstein M (2012) "Emerging themes in SecA2-mediated protein export", Nat. Rev. Microbiol. **10**(11): 779-789.

Fernandez-Tornero C, Lopez R, Garcia E, Gimenez-Gallego G, Romero A (2001) "A novel solenoid fold in the cell wall anchoring domain of the pneumococcal virulence factor LytA", Nat. Struct. Biol. **8**(12): 1020-1024.

Ferner-Ortner J, Mader C, Ilk N, Sleytr UB, Egelseer EM (2007) "High-affinity interaction between the S-layer protein SbsC and the secondary cell wall polymer of *Geobacillus stearothermophilus* ATCC 12980 determined by surface plasmon resonance technology", J. Bacteriol. **189**(19): 7154-7158.

Ferreira TG, Moura H, Barr JR, Pilotto Domingues RMC, Ferreira EO (2017) "Ribotypes associated with *Clostridium difficile* outbreaks in Brazil display distinct surface protein profiles", Anaerobe **45**: 120-128.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A (2016) "The Pfam protein families database: towards a more sustainable future", Nucleic Acids Res. **44**(D1): D279-D285.

Flaig R, Romano P, Hall D (2013) "Data-collection options at the macromolecular crystallography beamline I04 at Diamond Light Source", Acta Cryst. A **69**(a1): s32.

Fox T, de Miguel E, Mort JS, Storer AC (1992) "Potent slow-binding inhibition of cathepsin B by its propeptide", Biochem. **31**(50): 12571-12576.

Freeman J, Wilcox MH (2003) "The effects of storage conditions on viability of *Clostridium difficile* vegetative cells and spores and toxin activity in human faeces", J. Clin. Pathol. **56**(2): 126-128.

Fushinobu S, Alves VD, Coutinho PM (2013) "Multiple rewards from a treasure trove of novel glycoside hydrolase and polysaccharide lyase structures: new folds, mechanistic details, and evolutionary relationships", Curr. Opin. Struct. Biol. **23**(5): 652-659.

Ganeshapillai J, Vinogradov E, Rousseau J, Weese JS, Monteiro MA (2008) "*Clostridium difficile* cell-surface polysaccharides composed of pentaglycosyl and hexaglycosyl phosphate repeating units", Carbohydr. Res. **343**(4): 703-710.

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A (2003) "ExPASy: The proteomics server for in-depth protein knowledge and analysis", Nucleic Acids Res. **31**(13): 3784-3788.

Gessler F, Bohnel H (2006) "Persistence and mobility of a *Clostridium botulinum* spore population introduced to soil with spiked compost", FEMS Microbiol. Ecol. **58**(3): 384-393.

Gibbs GM, Roelants K, O'Bryan MK (2008) "The CAP superfamily: cysteine-rich secretory proteins, antigen 5, and pathogenesis-related 1 proteins--roles in reproduction, cancer, and immune defense", Endocr. Rev. **29**(7): 865-897.

Godoy AS, Camilo CM, Kadowaki MA, Muniz HD, Espirito Santo M, Murakami MT, Nascimento AS, Polikarpov I (2016) "Crystal structure of beta1→6-galactosidase from *Bifidobacterium bifidum* S17: trimeric architecture, molecular determinants of the enzymatic activity and its inhibition by alpha-galactose", FEBS J. **283**(22): 4097-4112.

Gooyit MD, Janda KD (2016) "Modulation of the Surface-Layer Protein of *Clostridium difficile* through Cwp84 Inhibition", ACS Infect. Dis. **2**(7): 465-470.

Grady LM, Michtavy J, Oliver DB (2012) "Characterization of the *Escherichia coli* SecA signal peptide-binding site", J. Bacteriol. **194**(2): 307-316.

Graham DR, Garnham CP, Fu Q, Robbins J, Van Eyk JE (2005) "Improvements in two-dimensional gel electrophoresis by utilizing a low cost "in-house" neutral pH sodium dodecyl sulfate-polyacrylamide gel electrophoresis system", Proteomics **5**(9): 2309-2314.

Grosdidier A, Zoete V, Michielin O (2011) "SwissDock, a protein-small molecule docking web service based on EADock DSS", Nucleic Acids Res. **39**(Web): W270-W277.

Guarner F, Malagelada JR (2003) "Gut flora in health and disease", The Lancet **361**(9356): 512-519.

Guex N, Peitsch MC (1997) "SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling", Electrophoresis **18**(15): 2714-2723.

Hall IC (1935) "Intestinal Flora in New-Born Infants", Am. J. Dis. Child. **49**(2): 390.

Hamburger ZA, Brown MS, Isberg RR, Bjorkman PJ (1999) "Crystal structure of invasin: a bacterial integrin-binding protein", Science **286**(5438): 291-295.

Hanada K, Tamai M, Yamagishi M, Ohmura S, Sawada J, Tanaka I (1978) "Isolation and Characterization of E-64, a New Thiol Protease Inhibitor", Agric. Biol. Chem. **42**(3): 523-528.

Harrison SC (2004) "Whither structural biology?", Nat. Struct. Mol. Biol. **11**(1): 12-15.

Heintzmann R, Ficz G (2013) "Breaking the resolution limit in light microscopy", Methods Cell Biol. **114**: 525-544.

Hendrickson WA, Lattman EE (1970) "Representation of phase probability distributions for simplified combination of independent phase information", Acta Cryst. B **26**(2): 136-143.

Hendrickson WA, Teeter MM (1981) "Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur", Nature **290**(5802): 107-113.

Hendrickson WA (1991) "Determination of macromolecular structures from anomalous diffraction of synchrotron radiation", Science **254**(5028): 51-58.

Hennequin C, Porcheray F, Waligora-Dupriet A, Collignon A, Barc M, Bourlioux P, Karjalainen T (2001) "GroEL (Hsp60) of *Clostridium difficile* is involved in cell adherence", Microbiol. **147**(Pt 1): 87-96.

Hennequin C, Janoir C, Barc MC, Collignon A, Karjalainen T (2003) "Identification and characterization of a fibronectin-binding protein from *Clostridium difficile*", Microbiol. **149**(Pt 10): 2779-2787.

Hennig M, Jansonius JN, Terwisscha van Scheltinga AC, Dijkstra BW, Schlesier B (1995) "Crystal structure of concanavalin B at 1.65 Å resolution. An "inactivated" chitinase from seeds of *Canavalia ensiformis*", J. Mol. Biol. **254**(2): 237-246.

Henrissat B (1991) "A classification of glycosyl hydrolases based on amino acid sequence similarities", Biochem. J. **280 (Pt 2)**(2): 309-316.

Herbold DR, Glaser L (1975) "Interaction of N-acetylmuramic acid L-alanine amidase with cell wall polymers", J. Biol. Chem. **250**(18): 7231-7238.

Holm L, Rosenstrom P (2010) "Dali server: conservation mapping in 3D", Nucleic Acids Res. **38**(Web): W545-W549.

Hook G, Jacobsen JS, Grabstein K, Kindy M, Hook V (2015) "Cathepsin B is a New Drug Target for Traumatic Brain Injury Therapeutics: Evidence for E64d as a Promising Lead Drug Candidate", Front. Neurol. **6**: 178.

Hope H (1988) "Cryocrystallography of biological macromolecules: a generally applicable method", Acta Cryst. B **44 (Pt 1)**: 22-26.

Horton HR, Moran LA, Scrimgeour KG, Perry MD, Rawn JD (2006) "Principles of Biochemistry", Pearson Education

Howell PL, Smith GD (1992) "Identification of Heavy-Atom Derivatives by Normal Probability Methods", J. Appl. Cryst. **25**(1): 81-86.

Incardona MF, Bourenkov GP, Levik K, Pieritz RA, Popov AN, Svensson O (2009) "EDNA: a framework for plugin-based applications applied to X-ray experiment online data analysis", J. Synch. Rad. **16**(Pt 6): 872-879.

Ito Y, Bae SC, Chuang LS (2015) "The RUNX family: developmental regulators in cancer", Nat. Rev. Cancer **15**(2): 81-95.

Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF (2001) "Protein flexibility predictions using graph theory", Proteins **44**(2): 150-165.

Janoir C, Grenery J, Savariau-Lacomme MP, Collignon A (2004) "[Characterization of an extracellular protease from *Clostridium difficile*]", Pathol. Biol. **52**(8): 444-449.

Janoir C, Pechine S, Grosdidier C, Collignon A (2007) "Cwp84, a surface-associated protein of *Clostridium difficile*, is a cysteine protease with degrading activity on extracellular matrix proteins", J. Bacteriol. **189**(20): 7174-7180.

Jaskolski M, Gilski M, Dauter Z, Wlodawer A (2007) "Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them?", Acta Cryst. D **63**(Pt 5): 611-620.

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S (2014) "InterProScan 5: genome-scale protein function classification", Bioinformatics **30**(9): 1236-1240.

Kabsch W, Sander C (1983) "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features", Biopolymers **22**(12): 2577-2637.

Kabsch W (2010) "*XDS*", Acta Cryst. D **66**(Pt 2): 125-132.

Kachrimanidou M, Malisiovas N (2011) "*Clostridium difficile* infection: a comprehensive review", Crit. Rev. Microbiol. **37**(3): 178-187.

Kamphuis IG, Kalk KH, Swarte MB, Drenth J (1984) "Structure of papain refined at 1.65 A resolution", J. Mol. Biol. **179**(2): 233-256.

Kantardjieff KA, Rupp B (2003) "Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals", Protein Sci. **12**(9): 1865-1871.

Karjalainen T, Waligora-Dupriet AJ, Cerquetti M, Spigaglia P, Maggioni A, Mauri P, Mastrantonio P (2001) "Molecular and genomic analysis of genes encoding surface-anchored proteins from *Clostridium difficile*", Infect. Immun. **69**(5): 3442-3446.

Karplus PA, Diederichs K (2012) "Linking crystallographic model and data quality", Science **336**(6084): 1030-1033.

Karplus PA, Diederichs K (2015) "Assessing and maximizing data quality in macromolecular crystallography", Curr. Opin. Struct. Biol. **34**: 60-68.

Kawata T, Takeoka A, Takumi K, Masuda K (1984) "Demonstration and Preliminary Characterization of a Regular Array in the Cell-Wall of *Clostridium difficile*", FEMS Microbiol. Lett. **24**(2-3): 323-328.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ (2015) "The Phyre2 web portal for protein modeling, prediction and analysis", Nat. Protoc. **10**(6): 845-858.

Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC (1958) "A three-dimensional model of the myoglobin molecule obtained by x-ray analysis", Nature **181**(4610): 662-666.

Kirby JM, Ahern H, Roberts AK, Kumar V, Freeman Z, Acharya KR, Shone CC (2009) "Cwp84, a surface-associated cysteine protease, plays a role in the maturation of the surface layer of *Clostridium difficile*", J. Biol. Chem. **284**(50): 34666-34673.

Kirby JM, Thiyagarajan N, Roberts AK, Shone CC, Acharya KR (2011) "Expression, purification, crystallization and preliminary crystallographic analysis of a putative *Clostridium difficile* surface protein Cwp19", Acta Cryst. F **67**(Pt 7): 762-767.

Kirby JM (2011) "The pathogenesis of *Clostridium difficile* infection", **PhD**, University of Bath, Bath, UK.

Kirk JA, Banerji O, Fagan RP (2017) "Characteristics of the *Clostridium difficile* cell envelope and its importance in therapeutics", Microbial biotechnology **10**(1): 76-90.

Kiyohara M, Nakatomi T, Kurihara S, Fushinobu S, Suzuki H, Tanaka T, Shoda S, Kitaoka M, Katayama T, Yamamoto K, Ashida H (2012) "alpha-N-acetylgalactosaminidase from infant-associated *Bifidobacteria* belonging to novel glycoside hydrolase family 129 is implicated in alternative mucin degradation pathway", J. Biol. Chem. **287**(1): 693-700.

Kovacs-Simon A, Leuzzi R, Kasendra M, Minton N, Titball RW, Michell SL (2014) "Lipoprotein CD0873 is a novel adhesin of *Clostridium difficile*", J. Infect. Dis. **210**(2): 274-284.

Langer G, Cohen SX, Lamzin VS, Perrakis A (2008) "Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7", Nat. Protoc. **3**(7): 1171-1179.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) "Clustal W and Clustal X version 2.0", Bioinformatics **23**(21): 2947-2948.

Lawley TD, Croucher NJ, Yu L, Clare S, Sebaihia M, Goulding D, Pickard DJ, Parkhill J, Choudhary J, Dougan G (2009) "Proteomic and genomic characterization of highly infectious *Clostridium difficile* 630 spores", J. Bacteriol. **191**(17): 5377-5386.

Lawson PA, Citron DM, Tyrrell KL, Finegold SM (2016) "Reclassification of *Clostridium difficile* as *Clostridioides difficile* (Hall and O'Toole 1935) Prevot 1938", Anaerobe **40**: 95-99.

Lazarevic V, Margot P, Soldo B, Karamata D (1992) "Sequencing and analysis of the *Bacillus subtilis* lytRABC divergon: a regulatory unit encompassing the structural genes of the N-acetylmuramoyl-L-alanine amidase and its modifier", J. Gen. Microbiol. **138**(9): 1949-1961.

Li YT, Cai HF, Wang ZH, Xu J, Fang JY (2016) "Systematic review with meta-analysis: long-term outcomes of faecal microbiota transplantation for *Clostridium difficile* infection", Aliment. Pharmacol. Ther. **43**(4): 445-457.

Liakopoulos A, Mevius D, Ceccarelli D (2016) "A Review of SHV Extended-Spectrum beta-Lactamases: Neglected Yet Ubiquitous", Front. Microbiol. **7**: 1374.

Little DJ, Li G, Ing C, DiFrancesco BR, Bamford NC, Robinson H, Nitz M, Pomes R, Howell PL (2014) "Modification and periplasmic translocation of the biofilm exopolysaccharide poly-beta-1,6-N-acetyl-D-glucosamine", Proc. Natl. Acad. Sci. **111**(30): 11013-11018.

Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2014) "The carbohydrate-active enzymes database (CAZy) in 2013", Nucleic Acids Res. **42**(Database): D490-D495.

Luo Y, Frey EA, Pfuetzner RA, Creagh AL, Knoechel DG, Haynes CA, Finlay BB, Strynadka NC (2000) "Crystal structure of enteropathogenic *Escherichia coli* intimin-receptor complex", Nature **405**(6790): 1073-1077.

Maffioli SI, Zhang Y, Degen D, Carzaniga T, Del Gatto G, Serina S, Monciardini P, Mazzetti C, Guglierame P, Candiani G, Chiriac AI, Facchetti G, Kaltofen P, Sahl HG, Deho G, Donadio S, Ebright RH (2017) "Antibacterial Nucleoside-Analog Inhibitor of Bacterial RNA Polymerase", Cell **169**(7): 1240-1248 e1223.

Manuel SG, Ragunath C, Sait HB, Izano EA, Kaplan JB, Ramasubbu N (2007) "Role of active-site residues of dispersin B, a biofilm-releasing beta-hexosaminidase from a periodontal pathogen, in substrate hydrolysis", FEBS J. **274**(22): 5987-5999.

Martin SHC (1885) "Papaïn-Digestion", J. Physiol. **5**(4-6): 213-230.

Matsumoto K, Mizoue K, Kitamura K, Tse WC, Huber CP, Ishida T (1999) "Structural basis of inhibition of cysteine proteases by E-64 and its derivatives", Biopolymers **51**(1): 99-107.

Matthews BW (1968) "Solvent content of protein crystals", J. Mol. Biol. **33**(2): 491-497.

Mauri PL, Pietta PG, Maggioni A, Cerquetti M, Sebastianelli A, Mastrantonio P (1999) "Characterization of surface layer proteins from *Clostridium difficile* by liquid chromatography/electrospray ionization mass spectrometry", Rapid Comm. Mass Spec. **13**(8): 695-703.

Mayer BJ (2001) "SH3 domains: complexity in moderation", J. Cell Sci. **114**(Pt 7): 1253-1263.

McCoubrey J, Poxton IR (2001) "Variation in the surface layer proteins of *Clostridium difficile*", FEMS Immunol. Med. Microbiol. **31**(2): 131-135.

McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ (2007) "Phaser crystallographic software", J. Appl. Cryst. **40**(Pt 4): 658-674.

McCoy AJ (2007) "Solving structures of protein complexes by molecular replacement with Phaser", Acta Cryst. D **63**(Pt 1): 32-41.

McFarland LV, Ozen M, Dinleyici EC, Goh S (2016) "Comparison of pediatric and adult antibiotic-associated diarrhea and *Clostridium difficile* infections", World Journal of Gastroenterology **22**(11): 3078-3104.

McGowan AP, Lalayiannis LC, Sarma JB, Marshall B, Martin KE, Welfare MR (2011) "Thirty-day mortality of *Clostridium difficile* infection in a UK National Health Service Foundation Trust between 2002 and 2008", J. Hosp. Infect. **77**(1): 11-15.

Merrigan MM, Venugopal A, Roxas JL, Anwar F, Mallozzi MJ, Roxas BA, Gerding DN, Viswanathan VK, Vedantam G (2013) "Surface-layer protein A (SlpA) is a major contributor to host-cell adherence of *Clostridium difficile*", PloS one **8**(11): e78404.

Miller WH (1839) "A treatise on crystallography", J. & J. J. Deighton, Cambridge,

Minichino A, Habash J, Raftery J, Helliwell JR (2003) "The properties of (2Fo - Fc) and (Fo - Fc) electron-density maps at medium-to-high resolutions", Acta Cryst. D **59**(Pt 5): 843-849.

Monot M, Boursaux-Eude C, Thibonnier M, Vallenet D, Moszer I, Medigue C, Martin-Verstraete I, Dupuy B (2011) "Reannotation of the genome sequence of *Clostridium difficile* strain 630", J. Med. Microbiol. **60**(Pt 8): 1193-1199.

Moore AD, Held A, Terrapon N, Weiner J, 3rd, Bornberg-Bauer E (2014) "DoMosaics: software for domain arrangement visualization and domain-centric analysis of proteins", Bioinformatics **30**(2): 282-283.

Moriarty NW, Tronrud DE, Adams PD, Karplus PA (2016) "A new default restraint library for the protein backbone in Phenix: a conformation-dependent geometry goes mainstream", Acta Cryst. D **72**(Pt 1): 176-179.

Morris AL, MacArthur MW, Hutchinson EG, Thornton JM (1992) "Stereochemical quality of protein structure coordinates", Proteins **12**(4): 345-364.

Motz H (1951) "Applications of the Radiation from Fast Electron Beams", J. of Appl. Phys. **22**(5): 527-535.

Mueller M, Wang M, Schulze-Briese C (2012) "Optimal fine phi-slicing for single-photon-counting pixel detectors", Acta Crystallogr D Biol Crystallogr **68**(Pt 1): 42-56.

Murshudov GN, Skubak P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, Winn MD, Long F, Vagin AA (2011) "REFMAC5 for the refinement of macromolecular crystal structures", Acta Cryst. D **67**(Pt 4): 355-367.

Musil D, Zucic D, Turk D, Engh RA, Mayr I, Huber R, Popovic T, Turk V, Towatari T, Katunuma N, et al. (1991) "The refined 2.15 A X-ray crystal structure of human liver cathepsin B: the structural basis for its specificity", EMBO J. **10**(9): 2321-2330.

Nagler DK, Zhang R, Tam W, Sulea T, Purisima EO, Menard R (1999) "Human cathepsin X: A cysteine protease with unique carboxypeptidase activity", Biochem. **38**(39): 12648-12654.

Naumoff DG (2011a) "GHL1-GHL15: New families of the hypothetical glycoside hydrolases", Mol. Biol. **45**(6): 983-992.

Naumoff DG (2011b) "Hierarchical classification of glycoside hydrolases", Biochem. (Moscow) **76**(6): 622-635.

Navaza J (1994) "AMoRe - an automated package for molecular replacement", Acta Cryst. A **50**(2): 157-163.

Ness SR, de Graaff RA, Abrahams JP, Pannu NS (2004) "CRANK: new methods for automated macromolecular crystal structure solution", Structure **12**(10): 1753-1761.

Pannu NS, McCoy AJ, Read RJ (2003) "Application of the complex multivariate normal distribution to crystallographic methods with insights into multiple isomorphous replacement phasing", Acta Cryst. D **59**(Pt 10): 1801-1808.

Pannu NS, Read RJ (2004) "The application of multivariate statistical techniques improves single-wavelength anomalous diffraction phasing", Acta Cryst. D **60**(Pt 1): 22-27.

Patterson AL (1935) "A Direct Method for the Determination of the Components of Interatomic Distances in Crystals" Z. Krist. Cryst. Mat. **90:** 517.

Payan F, Flatman R, Porciero S, Williamson G, Juge N, Roussel A (2003) "Structural analysis of xylanase inhibitor protein I (XIP-I), a proteinaceous xylanase inhibitor from wheat (*Triticum aestivum*, var. Soisson)", Biochem. J. **372**(Pt 2): 399-405.

Peltier J, Courtin P, El Meouche I, Lemee L, Chapot-Chartier MP, Pons JL (2011) "*Clostridium difficile* has an original peptidoglycan structure with a high level of N-acetylglucosamine deacetylation and mainly 3-3 cross-links", J. Biol. Chem. **286**(33): 29053-29062.

Peltier J, Shaw HA, Wren BW, Fairweather NF (2017) "Disparate subcellular location of putative sortase substrates in *Clostridium difficile*", Sci. Rep. **7**(1): 9204.

Perutz MF (1949) "An X-ray study of horse methemoglobin", Proc. R. Soc. A: Mat. Phys. Eng. Sci. **195**(1043): 474-499.

Perutz MF (1956) "Isomorphous Replacement and Phase Determination in Non-Centrosymmetric Space Groups", Acta Cryst. **9**(10): 867-873.

Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) "SignalP 4.0: discriminating signal peptides from transmembrane regions", Nat. Methods **8**(10): 785-786.

Pickersgill RW, Harris GW, Garman E (1992) "Structure of Monoclinic Papain at 1.60-a Resolution", Acta Cryst. B **48**(1): 59-67.

Podobnik M, Kuhelj R, Turk V, Turk D (1997) "Crystal structure of the wild-type human procathepsin B at 2.5 A resolution reveals the native active site of a papain-like cysteine protease zymogen", J. Mol. Biol. **271**(5): 774-788.

Postma N, Kiers D, Pickkers P (2015) "The challenge of *Clostridium difficile* infection: Overview of clinical manifestations, diagnostic tools and therapeutic options", Int. J. Antimicrob. Agents **46 Suppl 1**: S47-S50.

Pothineni SB, Venugopalan N, Ogata CM, Hilgart MC, Stepanov S, Sanishvili R, Becker M, Winter G, Sauter NK, Smith JL, Fischetti RF (2014) "Tightly integrated single- and multi-crystal data collection strategy calculation and parallelized data processing in JBluIce beamline control system", J. Appl. Cryst. **47**(Pt 6): 1992-1999.

Praznikar J, Turk D (2014) "Free kick instead of cross-validation in maximum-likelihood refinement of macromolecular crystal structures", Acta Cryst. D **70**(Pt 12): 3124-3134.

Qazi O, Hitchen P, Tissot B, Panico M, Morris HR, Dell A, Fairweather N (2009) "Mass spectrometric analysis of the S-layer proteins from *Clostridium difficile* demonstrates the absence of glycosylation", J. Mass Spec. **44**(3): 368-374.

Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) "Stereochemistry of polypeptide chain configurations", J. Mol. Biol. **7**: 95-99.

Raman R, Rajanikanth V, Palaniappan RU, Lin YP, He H, McDonough SP, Sharma Y, Chang YF (2010) "Big domains are novel Ca^{2+}-binding modules: evidences from big domains of *Leptospira* immunoglobulin-like (Lig) proteins", PloS one **5**(12): e14377.

Reddy VS, Natchiar SK, Stewart PL, Nemerow GR (2010) "Crystal structure of human adenovirus at 3.5 A resolution", Science **329**(5995): 1071-1075.

Redzynia I, Ljunggren A, Abrahamson M, Mort JS, Krupa JC, Jaskolski M, Bujacz G (2008) "Displacement of the occluding loop by the parasite protein, chagasin, results in efficient inhibition of human cathepsin B", J. Biol. Chem. **283**(33): 22815-22825.

Renko M, Pozgan U, Majera D, Turk D (2010) "Stefin A displaces the occluding loop of cathepsin B only by as much as required to bind to the active site cleft", FEBS J. **277**(20): 4338-4345.

Reynolds CB, Emerson JE, de la Riva L, Fagan RP, Fairweather NF (2011) "The *Clostridium difficile* cell wall protein CwpV is antigenically variable between strains, but exhibits conserved aggregation-promoting function", PLoS Path. **7**(4): e1002024.

Rhodes G (1993) "Crystallography made crystal clear : a guide for users of macromolecular models", Academic Press, San Diego

Rigel NW, Braunstein M (2008) "A new twist on an old pathway--accessory secretion systems", Mol. Microbiol. **69**(2): 291-302.

Rontgen WC (1896) "On a New Kind of Rays", Science **3**(59): 227-231.

Rossmann MG, Blow DM (1962) "The detection of sub-units within the crystallographic asymmetric unit", Acta Cryst. **15**(1): 24-31.

Rossmann MG (1990) "The molecular replacement method", Acta Cryst. A **46 (Pt 2)**: 73-82.

Rossmann MG (2001) "Molecular replacement--historical background", Acta Cryst. D **57**(Pt 10): 1360-1366.

Roy S, Choudhury D, Aich P, Dattagupta JK, Biswas S (2012) "The structure of a thermostable mutant of pro-papain reveals its activation mechanism", Acta Cryst. D **68**(Pt 12): 1591-1603.

Rupp B (2010) "Biomolecular crystallography : principles, practice, and application to structural biology", Garland Science, New York

Sajid M, McKerrow JH (2002) "Cysteine proteases of parasitic organisms", Mol. Biochem. Parasitol. **120**(1): 1-21.

Sara M, Sleytr UB (2000) "S-Layer proteins", J. Bacteriol. **182**(4): 859-868.

Savariau-Lacomme MP, Lebarbier C, Karjalainen T, Collignon A, Janoir C (2003) "Transcription and analysis of polymorphism in a cluster of genes encoding surface-associated proteins of *Clostridium difficile*", J. Bacteriol. **185**(15): 4461-4470.

Schaffer C, Messner P (2017) "Emerging facets of prokaryotic glycosylation", FEMS Microbiol. Rev. **41**(1): 49-91.

Schaschke N, Assfalg-Machleidt I, Machleidt W, Moroder L (1998) "Substrate/propeptide-derived endo-epoxysuccinyl peptides as highly potent and selective cathepsin B inhibitors", FEBS Lett. **421**(1): 80-82.

Schechter I, Berger A (1967) "On the size of the active site in proteases. I. Papain", Biochem. Biophys. Res. Commun. **27**(2): 157-162.

Scheres SH (2014) "Beam-induced motion correction for sub-megadalton cryo-EM particles", Elife **3**: e03665.

Schnablegger H, Singh Y (2013) "The SAXS Guide: Getting acquainted with the principles", Anton Paar GmbH, Austria

Scott RD (2009) "The Direct Medical costs of Healthcare-Associated Infections in U.S. Hospitals and the Benefits of Prevention". DoBq promotion, Centers for Disease Control and Prevention**:** 1-13.

Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, Stabler R, Thomson NR, Roberts AP, Cerdeno-Tarraga AM, Wang H, Holden MT, Wright A, Churcher C, Quail MA, Baker S, Bason N, Brooks K, Chillingworth T, Cronin A, Davis P, Dowd L, Fraser A, Feltwell T, Hance Z, Holroyd S, Jagels K, Moule S, Mungall K, Price C, Rabbinowitsch E, Sharp S, Simmonds M, Stevens K, Unwin L, Whithead S, Dupuy B, Dougan G, Barrell B, Parkhill J (2006) "The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome", Nat. Genet. **38**(7): 779-786.

Sekulovic O, Ospina Bedoya M, Fivian-Hughes AS, Fairweather NF, Fortier LC (2015) "The *Clostridium difficile* cell wall protein CwpV confers phase-variable phage resistance", Mol. Microbiol. **98**(2): 329-342.

Senzani S, Li D, Bhaskar A, Ealand C, Chang J, Rimal B, Liu C, Joon Kim S, Dhar N, Kana B (2017) "An Amidase_3 domain-containing N-acetylmuramyl-L-alanine amidase is required for mycobacterial cell division", Sci. Rep. **7**(1): 1140.

Shah DS, Joucla G, Remaud-Simeon M, Russell RR (2004) "Conserved repeat motifs and glucan binding by glucansucrases of oral *streptococci* and *Leuconostoc mesenteroides*", J. Bacteriol. **186**(24): 8301-8308.

Sharon N, Lis H (2004) "History of lectins: from hemagglutinins to biological recognition molecules", Glycobiol. **14**(11): 53R-62R.

Sheldrick GM (2008) "A short history of SHELX", Acta Cryst. A **64**(Pt 1): 112-122.

Shugar D (1952) "The measurement of lysozyme activity and the ultra-violet inactivation of lysozyme", Biochim. Biophys. Acta **8**(3): 302-309.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG (2011) "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega", Mol. Syst. Biol. **7**: 539.

Siklos M, BenAissa M, Thatcher GR (2015) "Cysteine proteases as therapeutic targets: does selectivity matter? A systematic review of calpain and cathepsin inhibitors", Acta Pharm. Sin. B **5**(6): 506-519.

Sivaraman J, Lalumiere M, Menard R, Cygler M (1999) "Crystal structure of wild-type human procathepsin K", Prot. Sci. **8**(2): 283-290.

Skubak P, Pannu NS (2013) "Automatic protein structure solution from weak X-ray data", Nat. Comms. **4**: 2777.

Sleytr UB, Beveridge TJ (1999) "Bacterial S-layers", Trends Microbiol. **7**(6): 253-260.

Smarda J, Smajs D, Komrska J, Krzyzanek V (2002) "S-layers on cell walls of cyanobacteria", Micron **33**(3): 257-277.

Somogyi A, Medjoubi K, Baranton G, Le Roux V, Ribbens M, Polack F, Philippot P, Samama JP (2015) "Optical design and multi-length-scale scanning spectro-microscopy possibilities at the Nanoscopium beamline of Synchrotron Soleil", J. Synch. Rad. **22**(4): 1118-1129.

Song Y, DiMaio F, Wang RY, Kim D, Miles C, Brunette T, Thompson J, Baker D (2013) "High-resolution comparative modeling with RosettaCM", Structure **21**(10): 1735-1742.

Spigaglia P, Galeotti CL, Barbanti F, Scarselli M, Van Broeck J, Mastrantonio P (2011) "The LMW surface-layer proteins of *Clostridium difficile* PCR ribotypes 027 and 001 share common immunogenic properties", J. Med. Microbiol. **60**(Pt 8): 1168-1173.

Stoka V, Turk V, Turk B (2016) "Lysosomal cathepsins and their regulation in aging and neurodegeneration", Ageing Res. Rev. **32**: 22-37.

Suhre K, Sanejouand YH (2004) "ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement", Nucleic Acids Res. **32**(Web): W610-W614.

Sumner JB, Howell SF (1936) "Identification of Hemagglutinin of Jack Bean with Concanavalin A", J. Bacteriol. **32**(2): 227-237.

Sutherland TE, Andersen OA, Betou M, Eggleston IM, Maizels RM, van Aalten D, Allen JE (2011) "Analyzing airway inflammation with chemical biology: dissection of acidic mammalian chitinase function with a selective drug-like inhibitor", Chem. Biol. **18**(5): 569-579.

Tartof K, Hobbs C (1987) "Improved media for growing plasmid and cosmid clones", Focus **9**(2): 12.

Tasteyre A, Barc MC, Collignon A, Boureau H, Karjalainen T (2001) "Role of FliC and FliD flagellar proteins of *Clostridium difficile* in adherence and gut colonization", Infect. Immun. **69**(12): 7937-7940.

Taylor G (2003) "The phase problem", Acta Cryst. D **59**(Pt 11): 1881-1890.

Thomas SM, Brugge JS (1997) "Cellular functions regulated by Src family kinases", Annu. Rev. Cell. Dev. Biol. **13**(1): 513-609.

Toh EC, Huq NL, Dashper SG, Reynolds EC (2010) "Cysteine protease inhibitors: from evolutionary relationships to modern chemotherapeutic design for the treatment of infectious diseases", Curr. Prot. Pept. Sci. **11**(8): 725-743.

Tsuji A, Kikuchi Y, Ogawa K, Saika H, Yuasa K, Nagahama M (2008) "Purification and characterization of cathepsin B-like cysteine protease from cotyledons of daikon radish, *Raphanus sativus*", FEBS J. **275**(21): 5429-5443.

Turk D, Podobnik M, Kuhelj R, Dolinar M, Turk V (1996) "Crystal structures of human procathepsin B at 3.2 and 3.3 Angstroms resolution reveal an interaction motif between a papain-like cysteine protease and its propeptide", FEBS Lett. **384**(3): 211-214.

Usenik A, Renko M, Mihelic M, Lindic N, Borisek J, Perdih A, Pretnar G, Muller U, Turk D (2017) "The CWB2 Cell Wall-Anchoring Module Is Revealed by the Crystal Structures of the *Clostridium difficile* Cell Wall Proteins Cwp8 and Cwp6", Structure **25**(3): 514-521.

Vagin A, Teplyakov A (1997) "MOLREP: an automated program for molecular replacement", J. Appl. Cryst. **30**(6): 1022-1025.

Vagin A, Teplyakov A (2010) "Molecular replacement with MOLREP", Acta Cryst. D **66**(Pt 1): 22-25.

van Wyk N, Drancourt M, Henrissat B, Kremer L (2017) "Current perspectives on the families of glycoside hydrolases of *Mycobacterium tuberculosis*: their importance and prospects for assigning function to unknowns", Glycobiol. **27**(2): 112-122.

Viars S, Valentine J, Hernick M (2014) "Structure and function of the LmbE-like superfamily", Biomolecules **4**(2): 527-545.

Viswanathan VK, Mallozzi MJ, Vedantam G (2010) "*Clostridium difficile* infection: An overview of the disease and its pathogenesis, epidemiology and interventions", Gut microbes **1**(4): 234-242.

Voth DE, Ballard JD (2005) "*Clostridium difficile* toxins: mechanism of action and role in disease", Clin. Microbiol. Rev. **18**(2): 247-263.

Walden H (2010) "Selenium incorporation using recombinant techniques", Acta Cryst. D **66**(Pt 4): 352-357.

Waligora AJ, Hennequin C, Mullany P, Bourlioux P, Collignon A, Karjalainen T (2001) "Characterization of a cell surface protein of *Clostridium difficile* with adhesive properties", Infect. Immun. **69**(4): 2144-2153.

Wang T, Zhang J, Zhang X, Xu C, Tu X (2013) "Solution structure of the Big domain from *Streptococcus pneumoniae* reveals a novel Ca^{2+}-binding module", Sci. Rep. **3**: 1079.

Waterman DG, Winter G, Gildea RJ, Parkhurst JM, Brewster AS, Sauter NK, Evans G (2016) "Diffraction-geometry refinement in the *DIALS* framework", Acta Cryst. D **72**(Pt 4): 558-575.

Weiss MS, Hilgenfeld R (1997) "On the use of the merging R factor as a quality indicator for X-ray data", J. Appl. Cryst. **30**(2): 203-205.

Weiss MS (2001) "Global indicators of X-ray data quality", J. Appl. Cryst. **34**(2): 130-135.

Wells S, Menor S, Hespenheide B, Thorpe MF (2005) "Constrained geometric simulation of diffusive motion in proteins", Phys. Biol. **2**(4): S127-S136.

Wen X, Yi LZ, Liu F, Wei JH, Xue Y (2016) "The role of cathepsin K in oral and maxillofacial disorders", Oral Dis. **22**(2): 109-115.

Weng Z, Rickles RJ, Feng S, Richard S, Shaw AS, Schreiber SL, Brugge JS (1995) "Structure-function analysis of SH3 domains: SH3 binding specificity altered by single amino acid substitutions", Mol. Cell Biol. **15**(10): 5627-5634.

Wiederanders B, Kaulmann G, Schilling K (2003) "Functions of propeptide parts in cysteine proteases", Curr. Prot. Pept. Sci. **4**(5): 309-326.

Wiederanders B (2003) "Structure-function relationships in class CA1 cysteine peptidase propeptides", Acta Biochim. Pol. **50**(3): 691-713.

Wiegand PN, Nathwani D, Wilcox MH, Stephens J, Shelbaya A, Haider S (2012) "Clinical and economic burden of *Clostridium difficile* infection in Europe: a systematic review of healthcare-facility-acquired infection", J. Hosp. Infect. **81**(1): 1-14.

Wierenga RK (2001) "The TIM-barrel fold: a versatile framework for efficient enzymes", FEBS Lett. **492**(3): 193-198.

Willing SE, Candela T, Shaw HA, Seager Z, Mesnage S, Fagan RP, Fairweather NF (2015) "*Clostridium difficile* surface proteins are anchored to the cell wall using CWB2 motifs that recognise the anionic polymer PSII", Mol. Microbiol. **96**(3): 596-608.

Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AG, McCoy A, McNicholas SJ, Murshudov GN, Pannu NS, Potterton EA, Powell HR, Read RJ, Vagin A, Wilson KS (2011) "Overview of the CCP4 suite and current developments", Acta Cryst. D **67**(Pt 4): 235-242.

Winter G (2010) "xia2: an expert system for macromolecular crystallography data reduction", J. Appl. Cryst. **43**(1): 186-190.

Winter G, Lobley CM, Prince SM (2013) "Decision making in xia2", Acta Cryst. D **69**(Pt 7): 1260-1273.

Wlodawer A, Minor W, Dauter Z, Jaskolski M (2008) "Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures", FEBS J. **275**(1): 1-21.

Wlodawer A (2017) Stereochemistry and Validation of Macromolecular Structures. Protein Crystallography: Methods and Protocols. A Wlodawer, Z Dauter, M Jaskolski. New York, NY, Springer New York**:** 595-610.

Wright A, Wait R, Begum S, Crossett B, Nagy J, Brown K, Fairweather N (2005) "Proteomic analysis of cell surface proteins from *Clostridium difficile*", Proteomics **5**(9): 2443-2452.

Wright A, Drudy D, Kyne L, Brown K, Fairweather NF (2008) "Immunoreactive cell wall proteins of *Clostridium difficile* identified by human sera", J. Med. Microbiol. **57**(Pt 6): 750-756.

Wurtz A, Bouchut E (1879) "[On the digestive ferment of *Carica papaya*]", Comptes rendus de l'Academie des Sciences **8**: 425-439.

Yamamoto Y, Kurata M, Watabe S, Murakami R, Takahashi SY (2002) "Novel cysteine proteinase inhibitors homologous to the proregions of cysteine proteinases", Curr. Prot. Pept. Sci. **3**(2): 231-238.

Yeats C, Rawlings ND, Bateman A (2004) "The PepSY domain: a regulator of peptidase activity in the microbial environment?", Trends Biochem. Sci. **29**(4): 169-172.

Yee AA, Savchenko A, Ignachenko A, Lukin J, Xu X, Skarina T, Evdokimova E, Liu CS, Semesi A, Guido V, Edwards AM, Arrowsmith CH (2005) "NMR and X-ray crystallography, complementary tools in structural proteomics of small proteins", J. Am. Chem. Soc. **127**(47): 16512-16517.

Yutin N, Galperin MY (2013) "A genomic update on clostridial phylogeny: Gram-negative spore formers and other misplaced clostridia", Environ. Microbiol. **15**(10): 2631-2641.

Zelensky AN, Gready JE (2005) "The C-type lectin-like domain superfamily", FEBS J. **272**(24): 6179-6217.

Zheng H, Hou J, Zimmerman MD, Wlodawer A, Minor W (2014) "The future of crystallography in drug discovery", Expert Opin. Drug Disc. **9**(2): 125-137.

Zhou Y, Ueda T, Muller M (2014) "Signal recognition particle and SecA cooperate during export of secretory proteins with highly hydrophobic signal sequences", PloS one **9**(4): e92994.

List of Publications

Bradshaw WJ, Kirby JM, Thiyagarajan N, Chambers CJ, Davies AH, Roberts AK, Shone CC, Acharya KR (2014) "The structure of the cysteine protease and lectin-like domains of Cwp84, a surface layer-associated protein from *Clostridium difficile*" Acta Cryst. D **70**(Pt 7): 1983-1993.

Bradshaw WJ, Roberts AK, Shone CC, Acharya KR (2015) "Cwp84, a *Clostridium difficile* cysteine protease, exhibits conformational flexibility in the absence of its propeptide" Acta Cryst. F **71**(Pt 3): 295-303.

Bradshaw WJ, Davies AH, Chambers CJ, Roberts AK, Shone CC, Acharya KR, (2015) “Molecular features of the sortase enzyme family”, FEBS J., 2096-2114

Bradshaw WJ, Kirby JM, Roberts AK, Shone CC, Acharya KR (2017a) "Cwp2 from *Clostridium difficile* exhibits an extended three domain fold and cell adhesion in vitro" FEBS J. **284**(17): 2886-2898.

Bradshaw WJ, Rehman S, Pham TT, Thiyagarajan N, Lee RL, Subramanian V, Acharya KR (2017b) "Structural insights into human angiogenin variants implicated in Parkinson's disease and Amyotrophic Lateral Sclerosis" Sci. Rep. **7**: 41996.

Bradshaw WJ, Kirby JM, Roberts AK, Shone CC, Acharya KR (2017c) "The molecular structure of the glycoside hydrolase domain of Cwp19 from *Clostridium difficile*" FEBS J. **284**(24): 4343-4357.

Bradshaw WJ, Roberts AK, Shone CC, Acharya KR (2017d) "The structure of the S-layer of *Clostridium difficile*" J. Cell Comm. Sig.